Detecting Cyberstalking from Social Media Platform(s) using Data Mining Analytics

A thesis submitted in partial fulfilment of the requirements of the University of West London for the degree of Doctor of Philosophy

by

Aimee Elizabeth Mirto

University of West London

Supervisors: Professor Shanyu Tang,

Professor Graham Brooks

January 2022

Abstract

Cybercrime is an increasing activity that leads to cyberstalking whilst making the use of data mining algorithms to detect or prevent cyberstalking from social media platforms imperative for this study. The aim of this study was to determine the prevalence of cyberstalking on the social media platforms using Twitter. To achieve the objective, machine learning models that perform data mining alongside the security metrics were used to detect cyberstalking from social media platforms.

The derived security metrics were used to flag up any suspicious cyberstalking content. Two datasets of detailed tweets were analysed using NVivo and R Programming. The dominant occurrence of cyberstalking was assessed with the induction of fifteen unigrams identified from the preliminary dataset such as “abuse”, “annoying”, “creep or creepy”, “fear”, “follow or followers”, “gender”, “harassment”, “messaging”, “relationships p/p”, “scared”, “stalker”, “technology”, “unwanted”, “victim”, and “violent”. Ordinal regression was used to analyse the use of the fifteen unigrams which were categorised according to degree or relationship/link towards cyberstalking on the platform Twitter.

Moreover, two lightweight machine learning algorithms were used for the model performance showcasing cyberstalking indicative content. K Nearest Neighbour and K Means Clustering were both coded in R computer language for the extraction, refined, analysation and visualisation process for this research. Results showed the emotional terms like “bad”, “sad” and “hate” were attached to the unigrams being linked to cyberstalking. Each emotional term was flagged up in correspondence with

one of the fifteen unigrams in tweets that correlate cyberstalking indicative content, proving one must accompany the other.

K Means Clustering results showed the two terms “bad” and “sad” were shown within 100 percent of the clustering results and the term “hate” was only seen within 60 percent of the results. Results also revealed that the accuracy of the KNN algorithm was up to 40% in predicting key terms-based cyberstalking content in a real Twitter dataset consisting of 1m data points.

This study emphasises the continuous relationship between the fifteen unigrams, emotional terms, and tweets within numerous datasets portrayed in this research, and reveals a general picture that cyberstalking indicative content in fact happens on Twitter at a vast rate with the corresponding links or relationships within the detection of cyberstalking.

Acknowledgements

For my mother, Marjorie.
In loving memory

First and foremost, I am extremely grateful to my supervisors, Professor Shanyu Tang and Professor Graham Brooks for their invaluable advice, continuous support, and patience during my PhD study. Their immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life. I would like to express my gratitude to Professor Shanyu Tang for his meticulous feedback and constant and continuous support. I am indebted to him, that despite his busy and tight schedule, for pushing me to perform to the best of my abilities throughout this process. I would like to thank all the members in the graduate programme. It is their kind help and support that have made my study and life in the UK a wonderful time.

I am thankful to my constantly supportive parents who have preserved in me a strong work ethics. Especially my father who has cheered me along from across the ocean. I am indebted to my friends near and abroad, without whom this journey would not have been possible. A special thank you to my PhD college and friend, Windy. Our afternoon walks with countless conversations about both of our journeys helped in a tremendous way, I thank you. Lastly, I particularly want to thank Michael for his countless love, support, and patience during this PhD course. When the pressures of work, family life and completing this degree got on top of me, he did whatever he could to lift the load off. Thank you for being by my side throughout this process.

Table of Contents

List of Figures

Chapter 1: Introduction

1.1 Overview

This thesis focuses on a potential development of detecting and preventing cyberstalking based on data mining analytics and machine learning that is used by the researcher. These metrics help detect and hypothetically prevent cyberstalking on the social media platform, Twitter. Moreover, with the use of reoccurring themes, unigrams, and tweets that illustrate how cyberstalking is materialising on Twitter. Further along, the structure of this thesis is outline in detail within this chapter, as well as what programmes were used and how those programmes were used for this research. In addition, this chapter mentioned the algorithms that the researcher used and how they help with the detection and possible prevention of cyberstalking on twitter, in connection with the use of unigrams and reoccurring themes. Lastly, a brief breakdown of what is to be expected in each chapter within this thesis.

Furthermore, the prompt growth of the never-ending internet along with the multiple chances to partake in cybercrime with the never-ending uses of the Internet or social media which pave the way for the cyberworld to evolve. The past decade has seen a rise in cybercrime due to the rapid development of the Internet. Research on hate crime sites and stalking are mostly conducted on what are restricted sites where membership is needed to access the material, postings and data online (Karyofyllis, 2018).

With the progression of cybercrime due to the Internet there are many branches or umbrella points of cybercrime and cyberstalking is one of those points and the main focus for this research. According to the Crown Prosecution Service (CPS), cyberstalking, i.e. harassment taking place on the Internet, includes “the use of social networking sites, chat rooms and other forums facilitated by technology”. (CPS, 2018)

“The internet can be used for a range of purposes relating to harassment, for example:

- to locate personal information about a victim;
- to communicate with the victim;
- as a means of surveillance of the victim;
- identity theft such as subscribing the victim to services, purchasing goods and services in their name;
- damaging the reputation of the victim;
- electronic sabotage such as spamming and sending viruses; or
- tricking other internet users into harassing or threatening a victim.” (CPS, 2018)

The importance to know how cyberstalking became, it is imperative to understand the origin term stalking. Originally, stalking involves behavioural invasion and referred to nonelectronic means of infringement. Stalking is related to a phenomenon referred to as obsessive relational intrusion (ORI). Again, (Sheldon et al., 2019) mention, ORI is an unwanted desire for intimacy through repetitive invasion of a person’s sense of physical or symbolic privacy. Alongside cyberstalking defined as the “act of stalking using the Internet, which can ultimately instigate threats, maltreatment, and/or harassment”, such acts occur on open platforms such as Twitter as much as membership only platforms / web sites. Twitter is used worldwide and is a platform in which users send, and read posts known as ‘tweets’ and interact (Rahman, Alotaibi, and Alsheri, 2019). Users post their opinions on places and people, but with Twitter it is possible to view ‘tweets’ without having an account via mainstream media. These ‘public’ tweets are sometimes abusive and threatening towards an individual or group. In this sense ‘stalking’ is not always on ‘closed’ platforms (e.g.,

Facebook where an account is needed or privatised). Therefore, it is possible to analyse 'tweets' without membership and with access only to platforms / web sites.

1.2 Background

Firstly, and most importantly without the Internet, all forms of cybercrime would not be apparent. Therefore, it is significant to understand why cybercrime is a vigorous form of crime and how it effects individuals around the world. As mentioned previously, the past decade has seen a rise in cybercrime due to the rapid development of the Internet. Research on hate crime sites and stalking are mostly conducted on what are restricted sites where membership is needed to access the material, postings and data online (Karyofyllis, 2018). However, cybercrime continues to rise and scale with complexity, even affecting essential services, businesses, and private individuals alike. Cybercrime costs the UK billions of pounds, causes untold damage, and threatens national security (National Crime Agency, web. accessed, 2021). Likewise, the Internet is a world-wide playground for many people and what they can do on that playground is astronomical, and that said playground is all at their fingertips.

Secondly, again without the Internet cybercrime would not have a major dependency, which can be a gateway for individuals to implement any other form of cybercrime or activities. For instance, again cyberstalking is a form of cybercrime that is impacted the ordinary criminal activity that is stalking. Cyberstalking defined as the "act of stalking using the Internet, which can ultimately instigate threats, maltreatment, and/or harassment". Several of these cyberstalking acts can occur on open platforms such as Twitter or any other microblogging site as well as the membership only platforms / web sites like Facebook or Instagram to name a few. Twitter which is a form of a microblogging site is used worldwide and is a platform in which users send, and read posts known as 'tweets' and interact (Rahman, Alotaibi, and Alsheri, 2019).

Users post their opinions on places and people, but with Twitter it is possible to view ‘tweets’ without having an account via mainstream media. These ‘public’ tweets are sometimes abusive and threatening. In this sense ‘stalking’ is not always on ‘closed’ platforms (e.g., Facebook where an account is needed). Therefore, it is possible to analyse ‘tweets’ without membership and with access only to platforms / web sites. Cyberstalking is a new and more advanced why to bring harm to an individual. With the Internet and the rapid growth of the never-ending network(s) that are social media, help pave the way for cyberstalking to take place and for the victims and their where abouts and information to be easily accessible. Sheldon, Rauschnabel, and Honeycutt, 2019 suggested: Cyberstalking is a serious predatory behaviour that arrives from the evolutionary need for control in the pursuit of resources and reputation.

A great example of a microblogging platform includes Twitter. Twitter is one of the best-known channels in the microblogging world. As stated before, Twitter is a quick and convenient way to share short posts, GIFs, article links, videos and more. Microblogging is effectively blogging done with severe space or size constraints typically by posting frequent brief messages about personal activities. Therefore, why the importance or the understanding of microblogging is the activity or practice of making short, frequent posts to a microblog. Because Twitter is the biggest and best-known or widely used microblogging platform and is the main platform that is the focus of this research.

On the other hand, numerous people have had their personal information hacked or stolen from another user on the Internet via social media or another connection networks. The number of large releases of personal information through hacking into IT systems over the last few years means that almost everyone who had

ever used a computer has had some aspect of their personal information stolen (Home Office, 2017). Social media has made sharing the majority of a person's personal information an art form; from sharing personal photos, to where people go for dinner, to where they travel, even to the names of their family members and friends, every second of their lives is documented online. Furthermore, the amount of legal information collected by marketers, social networks, and many other ways information is collected, sold, or shared through social networks is huge (Uzialko, 2018).

Furthermore, if hostile people want to know something about a person (the social media user), social media networks assist in increasing the accessibility of types of harassments, such as abuse and threats that are the hall marks of cyberstalking. Moreover, Soomro and Hussain (2019) illustrated that the number of Internet users has reached 4 168 461 500, i.e., 50.08 % penetration of world population and, in 2019 there were 2.77 billion social media networking users worldwide, i.e., 35.9 % of global social media networking penetration and it is expected that in 2021 this number will reach 3.02 billion.

Additionally, the Internet age has carried with it a number of tools and research which allows potential stalkers, either from ex-lovers, friends, acquaintances, to even complete strangers, to harass, threaten and abuse. Whilst, the majority of cyberstalking cases concern two (or more) ordinary people who were previously involved (Eterovic-Soric, Choo, Ashman, Mubarak, 2017). Stalking is a pattern of behaviour where: 'the legal definitions differ from country to country, even state to state (in the US), influenced by local stalking cases as laws were being enacted (similar to other cyber-related offences, such as online child exploitation Hillman, Hooper, Choo, 2014). This is where data analytics can be a useful aid to help prevent

and detect potential offenders/stalkers, as it is based on data, recording the volume of posting and abuse and velocity and frequency.

As seen throughout, the Internet age has conveyed roughly an unprecedented level of interconnectedness and the advances in communications technology have enabled friends and colleagues to keep in touch wherever they are in the world (Eterovic-Soric, et al, 2017). “Traditional crime typically occurs in one space and has an impact on one set of victims, whereas cybercrime can have a global impact” (United Kingdom 2010, 5). It is for the reasons and many more that cyberstalking can become vastly recognised and committed worldwide.

Besides, the current findings within academia, this research helps further current academic research on cyberstalking. As previously mentioned, academia and literature, mainly focus on cyber bullying, cyber fraud, and/or cyber hate crime. However, the term stalking has been mentioned throughout literature yet, with the use of the Internet and social media a new threat that is cyberstalking is emerging. In addition, the growth of social media and the Internet has made the term cyberstalking a new topic to be researched even further. Social media again, has made sharing the majority of a person’s personal information an art form. For example: from sharing personal photos, to where people go for dinner, to where they travel, even to the names of family and friends, every second of their lives is documented online. Finally, the importance of data mining and data analytics on cyberstalking would emphasise and/or help with the detection of origin of the cyberstalkers’ messages on any social media platform.

1.3 Research Rationale and Research Aim

All the previous studies of social media analytics (data mining-based) that are reported in the open literature focus on cyber fraud, cyber bullies, and cyber hate

crime. Cyberstalking analytics has not been given great attention by the researchers in the past and this motivated the present study. In addition, lightweight data mining algorithms have not been used to detect cyberstalking on social media platforms with the use of Twitter (Karyofyllis, 2018). The aim of this PhD research with the use of data mining and machine learning, is to have security metrics to detect cyberstalking from social media platforms with the use of Twitter.

In addition, as mentioned above and in the overview, the detection of cyberstalking, harassment, and security threats on Twitter by is undertaken by using the data mining analytics along with the algorithms and machine learning being used. The derived security metrics are then used to flag up any suspicious cyberstalking content (text-based and/or audio-based), to detect and prevent potential cyberstalking on social media platforms focusing on Twitter. With the expansion of the Internet, harassment, abuse, and threats increase in volume, velocity, and language. As such data mining and analytics can help detect the rise in the harassments and the threats that are a fragment of cyberstalking.

As previously mentioned, there is a gap within the literature within this field. Mentioned throughout this thesis, similarities to cyber bullying or cyber harassment are in comparison to cyberstalking. However, cyberstalking is not given the ample light or the noticeable recognising that is needed on this extensive topic. The conversation around cyberstalking with the use of this social media platform that was used for this thesis, is not happening within academia, unless cyber bullying or cyber harassment is attached to the topic at hand. Now, with it being known there is a narrative for topic cyberstalking, but not a stand-alone narrative, in connection with the social media platform, Twitter.

1.4 Research Questions

The four questions that the researcher is using for this study are as follows:

1. How can data mining and quantitative analysis of random open-sourced data samples reveal cyberstalking indicative content on social media platforms?
2. What security metrics indicate whether cyberstalking has been developed through social media platforms?
3. How can these metrics be used to provide a fine-grained measurement of cyberstalking?
4. Which data-mining algorithm is better suited for identifying and detecting cyberstalking on social media platforms?

1.5 Data Collection (Electronically and Manually): Preliminary Data

R-programming and NVivo were used to gather data on the topic and current research questions. These programmes consist of codes that will be mining data from secondary sources such as: literature-based texts, academic articles, journal articles, government websites, documents, and literature on cyberstalking. Both these programmes are student friendly and have free student license, however NVivo does have a paid preliminary service that is offered and has more tools to use.

1.5.1 NVivo and NCapture

NVivo, allows researchers to organise and analyse a wide variety of data, including but not limited to documents, images, audio, video, questionnaires, and web/ social media content (Edhlund and McDougall, 2019). The social media platform that is Twitter, was assessed by each programme. R-programming and NVivo are vastly new and forthcoming ways to conduct data analysis. They are extremely useful, record vast amount of data, help with analysing data, and as mentioned above a cost-effective method of research.

However, before the preliminary data collection began for the research, the two software programmes which were used during that process needed to be downloaded. The first software programme is NVivo with the extension that is offered with the programme which is called NCapture. The programme NVivo is used for regarding social media, it is one of the data mining techniques, which can be used to gather data on the topic and current research questions. NVivo is a software tool that complements the work of multi methodology research. NVivo is used for qualitative method as well as mixed method research. The NVivo programme consists of codes that mine data from secondary sources, such as literature-based texts, academic articles, journal articles, government websites, documents, and literature on cyberstalking. As mentioned before, NVivo is a software program used for qualitative and mixed-methods research. Specifically, it is used for the analysis of unstructured text, audio, video, and image data, including (but not limited to) interviews, focus groups, surveys, social media, and journal articles. Lastly, in addition as mentioned before the Google Chrome extension that is used with NVivo called NCapture, which was used for the process of taking threads of tweets from the social media platform Twitter.

NCapture was used for the collection of data that foraged the researcher on the preference of the unigrams that are the prime focus of the study. The researcher used NCapture to take thread of tweets from Twitter. Before, the researcher developed the unigrams that he or she would use for this study. He or she took to Twitter and did a few searches to see how the platform highlights certain aspects of cyberstalking. Furthermore, the researcher searched a few topics that have a strong reaction to cyberstalking and some topics that do not. Once, those topics were searched and tweets were captured with NCapture. The researcher then used the extension to gather all the materials from Twitter and import those materials into NVivo. Three

Twitter threads were used in the efforts to obtain the fifteen unigrams for this study. The three threads are: cyberstalking, #cybertalking, and stalking and fear. Each thread is talked about in detail within Chapter 3 of this thesis. After, the searches of Twitter threads were brought into NVivo, the researcher ran multiple word frequencies for each thread. Upon completion, the researcher then looked through the reoccurring themes and noticed a pattern within the word frequencies. Next, the researcher picked fifteen unigrams that he or she seemed adequate for the primal focus of this study. The unigrams are the centre of this study, other important avenues of this study focus on the fifteen unigrams that were found during the preliminary data collection process.

Formerly, once the researcher had the unigrams' he or she felt suited the study. He or she conducted Twitter searched of each unigram separately. He or she then used the Google Chrome extension, NCapture to capture 5,000 tweets in total for each unigram. The researcher imported the tweets back into NVivo from NCapture and when through each tweet one by one. Now, as he or she was manually looking through tweets for each unigram, he or she put the tweet(s) that had any correlation to cyberstalking in a Node. A node is a collection of references about a specific theme or case, for instance, in this study each unigram was a node and any tweet that referenced cyberstalking went into the corresponding node. The researcher did the same process for each node and along with each Twitter thread that was captured.

More so, the information that paralleled with each tweet as it was put into the node was also taken during this step. Immediately, after the codding for the node(s) process was done. The researcher counted how many of the tweets were: original tweets, re-tweets, and how many of the tweets used hashtags. This process was done again for each unigram, an original tweet is a tweet that someone posted from their account. A re-tweet is an original tweet that was reposted by completely different user. Lastly,

hashtags are added in a tweet for more clarity. Focus, or representation on the tweet, for example, if someone tweeted: I love dogs #goldenretriever. The hashtag that was used in the simple example, brings more clarity or representation to the tweet. The month the tweet was tweeted was also taken into consideration during the preliminary data collection. However, some of these characteristics were not used throughout this thesis but are a good steppingstone for furthering this research.

Immediately, after the data was collected and organised the researcher then exported the data from NVivo and into RStudio. The preliminary data was graphed within RStudio which supports and illuminates those findings even more. RStudio will be mentioned again in much detail with regard to the Random Sample Dataset that was used in chapter 4 of this thesis. While, using NVivo, this software programme helped the researcher analyse by allowing him or her to search for key themes, using a text or twitter search and a word frequency. Again, to narrow the amount to only those which evaluated preventive measures the research collected and ran multiple key word searches on Twitter, for the preliminary data collection process. These key words or terms were in comparison to cyberstalking to gather the material that was needed to conduct this research. Any Twitter thread or search that was included the following key unigrams were extracted:

- Abuse
- Annoying
- Creep/creepy
- Fear
- Follow/follows
- Gender
- Harassment

- Messaging
- Relationships P/P
- Scared
- Stalker
- Technology
- Unwanted
- Violent
- Victim

The preliminary data that was used in NVivo were the Twitter searches the researcher piloted for example: cyberstalking, stalking and sear, and #cyberstalking. During the analysis stage, the researcher used NVivo to carry out a qualitative text analysis which searched for the words and phrases used frequently in these searches from the Twitter threads. The preliminary data rational and reasoning was to highlight the importance and effectiveness of these unigrams that are considered to be beneficial in detecting and preventing cyberstalking on the social media platform that is Twitter. By using NVivo, the researcher had collated the tweets or threads from each unigram, in relation to themes that had previously emerged when conducting the original searches on Twitter.

1.6 Data Collection (Electronically): Random Sample Dataset

R (programming) is a programming language, which is used as a platform independent so it is compatible with any other operating system; using R the researcher can create objects, functions, and packages to analyse the data that is gathered (Dataflair; Web, 2019).

1.6.1 R Programme Language and RStudio

As mentioned beforehand, RStudio was used for the analysation process of the datasets provided in this study. RStudio is used for data mining techniques and is becoming more of an academic tool that is being used within academia. RStudio is best used with R. Simply, R is a programming language used for statistical computing while RStudio uses the R language to develop statistical programs. In R, people can write a program and run the code independently of any other computer program. However, R may be used without RStudio, but RStudio may not be used without R. The advantages of R Programming are endless and are very forthcoming as to why this programme is in fact very profitable to use for research purposes. Below are the various benefits of R language, which help grasp the concept of why it is so beneficial:

- **Open Source*:*** R is an open-source programming language. This means that anyone can work with R*,* without any need for a license or a fee. Furthermore, people can contribute towards the development of R by *customizing its packages, developing new ones and resolving issues.*

- **Exemplary Support for Data Wrangling*:*** R provides exemplary support for data wrangling. The packages like *dplyr, readr*are capable of transforming messy data into a structured form.

- **The Array of Packages*:*** R has a vast array of packages. With over 10,000 packages in the CRANrepository, the number is constantly growing. These packages appeal to all the areas of industry.

- **Quality Plotting and Graphing*:*** R facilitates quality plotting and graphing. The popular libraries like *ggplot2* and *plotly* advocate for aesthetic and visually appealing graphs that set R apart from other programming languages.

- **Highly Compatible:** R is highly compatible and can be paired with many other programming languages like C, C++, Java, and Python. It can also be integrated with technologies like Hadoop and various other database management systems as well.

- **Platform Independent:** R is a platform-independent language. It is a cross-platform programming language, meaning that it can be run quite easily on Windows, Linux, and Mac.

- **Eye-Catching Reports:** With packages like Shiny and Markdown, reporting the results of an analysis is extremely easy with R. People can make reports with the data, plots and R scripts embedded in them. People can even make interactive web apps that allow the user to play with the results and the data.

- **Machine Learning Operations:** R provides various facilities for carrying out machine learning operations like *classification, regression and also provides features for developing artificial neural networks.*

- **Statistics:** R is prominently known as the lingua franca of statistics. This is the main reason as to why R is dominant among other programming languages for developing statistical tools.

- **Continuously Growing:** R is a constantly evolving programming language. It is a state-of-the-art technology that provides updates whenever any new feature is added (Data-Flair, web, 2021).

As previously mentioned, and introduce in Chapter 4, a random sample dataset that was used for this study. This dataset had vast amounts of everyday tweets from everyday people. Likewise, there are five datasets with 50,000 tweets each in each saved csv file. The datasets are saved as a csv file for the purpose of importing them into RStudio. RStudio was used with each dataset to clean the dataset, strip the dataset of not needed information, as well as clean the dataset from whitespaces and numeric content, also, to remove any URLs, to remove the re-tweeted tweets and duplicate tweets. Once, that portion was finished the next step in RStudio was to make a term document matrix and then find the word frequencies for each dataset. After, each coding step was completed the graphing and word frequency was done. As well

as the researcher took the word frequency from RStudio and imported that information and ranked it for another visualisation of the data that was being analysed.

RStudio is also used in Chapter 5 with the tenacity of the use of algorithms to answer the research question number four: which data-mining algorithm is better suited for identifying and detecting cyberstalking on social media platforms? The algorithms that were used is K-means and K nearest neighbour. The researcher thought it would be favourable to use two algorithms, because with the results of the word frequencies from the random sample dataset in chapter four. That the results of the k-means algorithm would be ideally the same. Therefore, K nearest neighbour would be just as suitable with regards to the unigrams and the correspondence of cyberstalking.

1.7 Research Methods

1.7.1 Cluster analysis

Clustering is one of the most widely used forms of unsupervised learning. It's a great tool for making sense of unlabelled data and for grouping data into similar groups. A powerful clustering algorithm can decipher structure and patterns in a data set that are not apparent to the human eye. Therefore, clustering is the process of grouping the observed data into clusters based on some similarity or distance measure and then identifying subsequent data as belonging to a cluster. Moreover, cluster analysis is used as or for a statistical method for processing data. The reasoning why clustering works is by organising items into groups, or clusters, on the basis of how closely associated they are. In addition, as stated previously cluster analysis is an unsupervised learning algorithm, meaning that the people do not know how many clusters exist in the data before running the model. Unlike many other statistical methods, cluster analysis is typically used when there is no assumption made about

the likely relationships within the data. It provides information about where associations and patterns in data exist, but not what those might be or what they mean.

As stated above since clustering or cluster analysis is an unsupervised learning problem. It is often used as a data analysis technique for discovering interesting patterns in data, such as reoccurring themes that are text based, based on the current tested and its behaviours. Therefore, K means clustering algorithm was used within this study and is talked about and mentioned in detail further along. Furthermore, unsupervised, or undirected data science uncovers hidden patterns in unlabelled data. In unsupervised data science, there are no output variables to predict. Again, the main objective of this class or the analysis of data science techniques, is to find patterns in data based on the relationship between the data points themselves. More importantly, unsupervised learning is very useful in exploratory analysis because it can automatically identify structure in data. As well as, dimensionality reduction, which refers to the methods used to represent data using less columns or features, can be accomplished through unsupervised methods

Nevertheless, there are many clustering algorithms to choose from and no single best clustering algorithm for all cases. The reasoning as to why clustering is commonly used is because clustering in data mining helps in the classification of the data along with the certain datasets that are being used using similar functions or genes in the field of the present study. It helps in gaining insight into the structure of the research that it is accessing as well as many areas within that research are identified using the clustering in data mining. After data collection from the open twitter platform on the Internet, R-programming and NVivo were used in the PhD research to perform cluster analysis on the data collected.

1.7.2 Abnormal security patterns detection

Abnormal security patterns help provide total protection against the widest range of attacks including phishing, malware, ransomware, social engineering, executive impersonation, supply chain compromise, internal account compromise, spam, and graymail, and much more. The objective of the proposed research is to predict and detect cyberstalking from open-source social media data using lightweight data mining analytical algorithms against the following security metrics:

- Level of confidentiality
- Measure of integrity
- Degree of availability

To achieve the objective, the research needs to conduct predictions about abnormal patterns in social media data available on the Internet using machine learning data mining algorithms such as

- **K Means**: K Means Clustering Algorithm
- **KNN**: K Nearest Neighbour Algorithm

The importance of knowing the difference between unsupervised and supervised learning is imperative. Therefore, the main distinction concerning the two approaches is the use of labelled datasets. To put it simply, supervised learning uses labelled input and output data, while an unsupervised learning algorithm does not. Likewise, unsupervised learning models, in contrast, work on their own to discover the inherent structure of unlabelled data. Unsupervised learning is a machine learning technique, where people do not need to supervise the model. Supervised learning allows people to collect data or produce a data output from the previous experience. Unsupervised machine learning helps people to finds all kind of unknown patterns in data.

1.7.3 K Means Clustering

K means clustering in R Programming is an unsupervised non-linear algorithm that clusters data based on similarity or similar groups. Henceforth, why it was used in this study with the present datasets. Furthermore, k means has been around since the 1970s and fares better than other clustering algorithms such like density based and expectation maximisation. Moreover, this algorithm is seen as one of the most robust methods, especially for image segmentation and image annotation projects. However, according to some users, k means is very simple and easy to implement. Additionally, k means seeks to partition the observations into a pre-specified number of clusters. Segmentation of data takes place to assign each training example to a segment called a cluster. An advantage of k means is guaranteeing convergence. This algorithm can warm start the positions of centroids. As well as easily adapts to new examples and generalizes to clusters of different shapes and sizes, suchs as elliptical clusters.

K means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. Likewise, k means defines a target number k, which refers to the number of centroids that is needed in the dataset. A centroid is the imaginary or real location representing the centre of the cluster. After every point has been assigned, the centroid is moved to the average of all the points assigned to it. The algorithm is done when no point changes assigned centroid. Every data point is allocated to each of the clusters through reducing the in-cluster sum.

As mentioned, k means is an innovative algorithm that groups similar data into clusters. It calculates the centroids of k clusters and assigns a data point to that cluster

having least distance between its centroid and the data point. Here's how it works: start by choosing a value of k. for example, use k = 3. Then, randomly assign each data point to any of the 3 clusters. K means clustering is extensively used in various fields such as text mining, machine learning, image analysis, image processing, web cluster engines, bioinformatics, weather report, and so on (Bijuraj, 2013). Hence, why k means is being used for this study. It has been shown that k means is used in various fields and two of those fields are the main methods of this research: text mining and machine learning. Making k means the seamless algorithm to carry out this study along with the methods that the researcher has selected for this study. Moreover, there are diverse methods of the clustering for instance, model-based method, density-based method, hierarchical method, grid-based method, partitioned method.

Here is a breakdown of some the advantaged and the disadvantages of K-Means.

K -Means Advantages:

- If variables are huge, then K-Means most of the times computationally faster than hierarchical clustering, if k is kept small.
- K-Means produce tighter clusters than hierarchical clustering, especially if the clusters are globular.

K-Means Disadvantages:

- Difficult to predict K-Value.
- With global cluster, it does not work well.
- Different initial partitions can result in different final clusters.
- It does not work well with clusters (in the original data) which have different size and different density.

1.7.4 KNN: K-Nearest Neighbour

The abbreviation KNN stands for “K-Nearest Neighbour”. It is a supervised machine learning algorithm unlike the K-Means machine learning algorithm. The algorithm can be used to solve both classification and regression problem statements. The number of nearest neighbours to a new unknown variable that must be predicted or classified is denoted by the symbol 'K'. KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression). Determine parameter K = number of nearest neighbours. To use KNN it must calculate the distance between the query instance and all the training samples. Sort the distance and determine nearest neighbours based on the K, the minimum distance. Gather the category of the nearest neighbours.

The KNN algorithm can compete with the most accurate models because it makes highly accurate predictions. Therefore, the KNN algorithm can be used for applications that require high accuracy but that do not require a human readable model. However, the quality of the predictions within the model depends on the distance measure. KNN doesn’t learn any model that has to be done manually and was done for this study that is seen and mentioned more in Chapter 6 and Chapter 7 of this thesis. Lastly, KNN makes predictions using the similarity between an input sample and each training instance.

However, using KNN in RStudio is majorly important for this research. The knn () function needs to be used to train a model for which we need to install a package 'class'. The knn() function identifies the k-nearest neighbours using Euclidean distance where k is a user-specified number. Moreover, for the function “knn” in R. If it isn't

already installed, then the correct package must be installed. Likewise, to install the package, which as stated above is “knn” which will be present in RStudio library. Another beneficial package to use for KNN is also the package “caret”, this package is used within this study is talked in more detail within Chapter 6. The package “caret” contains different functions (as knn) for modelling complex regression and classification problems. On the other hand, while using the packages for KNN in RStudio, the process of cross-validation is just as important for the KNN model. Cross validation can be briefly described in the following steps: divide the data into k equally distributed chunks/folds. Then choose 1 chunk/fold as a test set and the rest K-1 as a training set. After that is done develop a KNN model based on the training set. Lastly, compare the predicted value vs actual values on the test set only.

Correspondingly, for a quick summary of KNN Algorithm would be k is a positive integer and with a new sample, it must specify k. Therefore, k is selected from database closest to the new sample. KNN works on a principle assuming every data point falling in near to each other is falling in the same class. In other words, it classifies a new data point based on similarity. Here are some advantages of KNN:

- Quick calculation time
- Simple algorithm – to interpret
- Versatile – useful for regression and classification
- High accuracy – it does not need to compare with better-supervised learning models
- No assumptions about data – no need to make additional assumptions, tune several parameters, or build a model. This makes it crucial in nonlinear data case.

As well as some of the disadvantages of using KNN**:**

- Accuracy depends on the quality of the data
- With large data, the prediction stage might be slow
- Sensitive to the scale of the data and irrelevant features
- Require high memory need to store all the training data
- Given that it stores all the training, it can be computationally expensive

1.8 Algorithm performance and efficiency measurements

The research was conducted by optimisation of prediction algorithms to identify the best way of detecting cyberstalking from open-source social media data. Comparisons between these data mining algorithms can be carried out in terms of algorithm complexity, level of security, computational overhead, and performance (detection accuracy, detection rate, and false alarm rate), to optimise the performance of the algorithms. Algorithm complexity measurement or computational overhead measurement: Charlier, Musumbu, and Hentenryck, 1991, explain: “A generic abstract interpretation algorithm is an algorithm, independent of the abstract domain that can be instantiated to provide an algorithm tailored to a specific application. Generic abstract interpretation algorithms are parametrized on the abstract domains in terms of abstract substitutions and a number of operations involving them. The abstract operations are consistent abstractions of the concrete operations in terms of which the concrete semantics is defined. Instantiating the generic abstract interpretation algorithm amounts to designing abstract substitutions capturing the relevant information and implementing a consistent operation on these substitutions”.

In addition, after many unsuccessful attempts, algorithm designers inevitably start to wonder, if there is something inherent in the problem that makes it impossible to devise algorithms that are faster than the current one. They may try to develop

mathematical techniques for proving formally that there can be no algorithm for the given problem which runs faster than the current one. Such a proof would be valuable, as it would suggest that it is futile to keep working on improved algorithms for this problem, that further improvements are certainly impossible. The realm of mathematical models and techniques for establishing such impossibility proofs is called computational complexity (Wiley and Stons Ldt., 2012). Level of security estimation: The basis for estimating data security level in sensor networks is to assess both the vulnerabilities and strengths of the different security mechanisms installed, such as key management schemes (KMSs) and intrusion detection systems (IDSs). This means that the impact of an attack in a sensor network depends on the effectiveness and vulnerabilities of the security mechanisms deployed. Quantifying the probability of the resistance of security mechanisms against attacks is exactly the goal of security level estimation activity (Ramos and Filho, 2015).

Ramos and Filho (2015) also mentioned, a network that has security mechanisms, but that does not have a system to estimate the security level provided by these mechanisms, may lead users to have a false sense of security. This happens because of the simple fact that a network has security mechanisms, but this does not necessarily imply that this network will be totally safe, exactly due to the vulnerabilities of these mechanisms. Unlike sensor networks, in traditional networks, there is a considerable amount of research and availability of standardised techniques for measuring security based on the evaluation of security mechanisms and on the analysis of network vulnerabilities, such as the Common Vulnerability Scoring System (CVSS) standard, which is used to quantify the severity of vulnerabilities (Ramos and Filho, 2015).

1.8.1 Performance measurement:

There are five specific types of measures that have been identified or defined and can be applied on performance measurement the are: input, output, efficiency, quality, and outcome. Within this study or research experiments were conducted on the research at hand to compare the two data mining algorithms in terms of detection accuracy and detection rate. Also, to simulate a practical environment, in the experiments, the algorithms can be run on the client and the server installed on two PCs over a local area network. The client is Intel Pentium G630 processor which offers a maximum clock speed of 2.7 GHz and 4GB memory. The hardware platform for the server is Intel Core (TM) i5 with 8GB memory which offers a maximum clock speed of 2.53 GHz.

The performance measurement is carried out when various algorithms are used to detect the abnormal security patterns and compared against each other. For example, Zhang, Tang, Cai, (2014): test their protocol and then compare it with relevant protocols in terms of computational cost. While comparing the protocols the researchers reported that their protocol was more informative in terms of withstanding replay attacks, impersonation attack, stolen-verifier attacks, and man-in-the-middle attacks.

The reasoning as to these methods were done throughout this thesis, because the researcher wanted to get an understanding of how the topic cyberstalking is being progressed on Twitter and how people's experiences are. The main reasoning as to why the research was conducted and designed this way, the researcher's aim was to identify the weighting of terms or unigrams in correlation towards cyberstalking on Twitter.

Sentiment analysis, known also as opinion mining was used, for text mining, classifying specific words into positive or negative (Rahman, AlOtaibi, Alsheri, 2019). In addition, both programmes (R-programming and NVivo) input graphs, charts, algorithms' word-art, pictures, or illustrations to support and exemplify the data analysis. Sentiment Analysis is a procedure used to determine if a chunk of text is positive, negative, or neutral. In text analytics, natural language processing (NLP) and machine learning (ML) techniques are combined to assign sentiment scores to the topics, categories, or entities within a phrase. These artificially intelligent bots are trained on millions of pieces of text to detect if a message is positive, negative, or neutral. Sentiment analysis works by breaking a message down into topic chunks and then assigning a sentiment score to each topic. For instance, the steps for sentiment analysis are data collection This is one of the most important steps in the sentiment analysis process. Data processing the processing of the data will depend on the kind of information it has either text, image, video, or audio. Lastly, data analysis and data visualization.

Furthermore, with NVivo and R-programming data that already is in the public sphere is an unlimited tool to use. Whilst the focus is on postings on Twitter this approach can include newspapers, novels, radio, the Internet, or archived data; these can provide valuable learning experiences as data is gathered from each source and analysed within the programme (Jackson and Bazeley, 2019). Pairing these two innovative and unique programmes together will benefit the improving narration of cyberstalking within academia.

The intention or reasoning as to why the use of data mining techniques on social media is that the data is enabling factor for advanced search in search engines

such as Twitter, and also helps in better understanding of data for research and organisational functions (Aggarwal, 2011).

Data mining techniques are capable of handling the three dominant disputes with social media, which are: size, noise, and dynamism. Each dispute is measured as in how the dispute is classified within the quantity of its mention or public perception. Therefore, social media data sets are very voluminous and require automated information processing for analysing it within a reasonable timeframe. Likewise, Adedoyin-Olowe, Gaber, and Stahl (2014) suggests: 'SM (social media) sites appear to be perfect sites to work on especially where opinion/sentiment expression is involved. Again, social media data sets are characterised by the three dominant disputes. These disputes are size, noise, and dynamism. Analysing these disputes and with the use of data mining on social media will classify the size or volume, noise of postings, and the dynamic of the 'tweets'. Aldedoyin-Olowe et al. (2014) explained: 'SM is characterised by noisy data such as spam blogs and irrelevant tweets in the case of Twitter. The dynamism in SM data sets are versatile in handling such dynamic data'.

The data mining method that is being used is to measure parameters such as:

- Terms / key words
- Number of postings / conversation or connections
- Probabilities (key words appear)
- Weightings of terms or key words
- Location of postings or connections (IP address)

1.9 Ethics

It is anticipated that ethical risks were low for this research as all data to be gathered is openly available in the public domain. There was no person-to-person contact and the main component of this part of the study is electronic. In addition, no person-to-person interviews or contact is required in the proposed research as it is about data mining and analytics. In addition, tweets were anonymous in the coding framework, and coding was done on secondary sources such as: literature, journals, blogs, scholarly articles, and social media and the Internet (all public accessed) using the researcher's own computer as well. There was no other individuals who will have access to the computer as well as the files that are being coded within the programmes themselves.

Furthermore, by using Twitter, which is a publicly accessible social media platform, a consent form is not needed. Twitter has evolved to be a credible medium of sentiments/opinion and expressions (Adedoyin-Olowe et al, 2014). Since the open social media network is used as an everyday tool that can be accessed by anyone, it is seen as public or communal knowledge.

1.10 Impact Statement

This research was intended to achieve an extensive understanding of the research questions being asked. This research aims were focus on, identifying cyberstalking, using a data mining method to identify metrics of cyberstalking. Therefore, it helps qualify a metrics measurement to determine the threshold value, either with communication or rapport and the linking of different types of patterns from normal online users compared to a potential online cyberstalking user(s). Lastly, this research can apply the proposed model to a real data set on social media.

Furthermore, the data analytics can determine whether there is a correlation between cyberstalking and the use of digital technology within social media networks.

This research that was conducted can help further current academic research on cyberstalking. In conclusion, data gathered within this study can inform and pave the way for a new area of communication with the use of recurring themes, negative and positive terms, from within social media network. Finally, the methodology and research conducted can advance current academic research in this field and continue to form an in-depth narration or conversation concerning the profound topic that is cyberstalking.

1.11 Layout of the thesis

The research described in the following chapters is set out in the following way: In chapter two, the literature review is given a perplex view on the topic at hand. In chapter three, the research methodology is described, the PhD candidate as researcher is declared, the preliminary data is collected and established. In chapter four, a random sample dataset is brought into the research scop to be compared with the preliminary data. Lastly, multiple methods of data collection and analysis are outlined throughout the thesis.

Chapter 1: sets the scene, this chapter is the introduction of the thesis and what is to come within the thesis: such as the research questions, what methods are being used, and lastly, how they were used. this chapter gives insight into the next chapters and furthers along the structure within this thesis.

Chapter 2: presents the literature review of past, present, and fairly recent literature. The chapter breakdown the important key points that help illustrate what cyberstalking

is, along with the background information that is needed. The literature review chapter starts off with the understanding of what cybercrime is and how it is the umbrella term that harvests cyberstalking. While then moving on to another supportive term, cybersecurity and the reasons as to why it is important to know and understand. Cybersecurity which informs the reader of the ins and outs of how to protect any cyber data that are stored on the Internet. Some interesting terms are highlighted within that section like, phishing and hacking to help develop the readers comprehension on cybersecurity even further.

Chapter 3: reviews the preliminary data that was taken from Twitter and imported into NVivo. Where word frequencies were run in NVivo on the Twitter threads obtained. Moreover, how the preliminary data was collected, searched on Twitter, and analysed. As well as, how each unigram was presented and collected and used for this study. Also, the 5,000 tweets per unigram were collected and analysed in this chapter. The chapter focuses on the importance of each unigram and why they were selected.

Chapter 4: begins with the preliminary dataset and its results. Next, the random sample dataset that was used for this study. Each dataset csv file, there are five in total. Each file was imported into RStudio and analysed. While in R each dataset was cleaned and analysed even further for this study. This chapter also, mentions how each unigram was searched within RStudio in correlation to cyberstalking.

Chapter 5: within this chapter the algorithm K Means Clustering is used on each csv file of the random sample dataset containing over 1 million data points. K means is used to detect cyberstalking content as well as the similar patterns of cyberstalking

within each dataset. The findings and analysis done within this chapter help assistance the research topic and its aims at hand. Such as the reoccurring themes being seen throughout the thesis or showcasing the cyberstalking indicative content.

Chapter 6: this chapter is a continuation of the above Chapter 5 K means results, however while using a different method/algorithm. Likewise, in this chapter a second algorithm is used for this study for instance, within this chapter K Nearest Neighbour (KNN) is used for the clustering results brought forth from Chapter 5. While KNN is the main method being used for the clustering results. In addition, a KNN performance model was made to be used on the clustering results to bring forth the results of the dataset taken from Chapter 5. Moreover, whilst the use of the KNN algorithm within this chapter is also compared to its previous method K Means Clustering.

Chapter 7: lastly, this chapter is the discussion and conclusion of the work at hand within this thesis. Such as answering all the research questions with the selected chapters or even more than one chapter that corroborates each question. As well as, any future works, limitations that this study had or prevented any work from happening. Also, mentioned are the recommendations towards this study and how they could benefit the outcome of the study. Which is then followed by the appendix and references for the entire thesis.

Chapter 2: Literature Review

2.1. Cybercrime

Cybercrime is defined as crime or illegal activity that is done using the Internet or network systems. Dr. Emma Ogilvie (2000), mentions:

"Cybercrime" has increasingly moved to the foreground in examinations of 21st century criminality. Arguments as to whether Internet-based technologies have created entirely new types of crime requiring equally new legislative and other responses, or simply provided for new expressions of traditional crimes requiring the adaptation of current legislative strategies, are hotly debated by the proponents of both views. At the heart of this dispute is the more important question of whether or not it is possible to regulate the Internet to anything like the same extent other communication media are controlled.

This research will narrow down this expansive topic of cybercrime, current literature lacks the narration of the linking or rapport on cyberstalking and the mainstream use of digital technology and/ or social media networks. A recent view of the literature recommends many definitions of cybercrime. Cybercrime can be defined as "criminal activities carried out by means of computers or the Internet" (Webster, 1995). However, an examination of cybercrime ought to begin with the Internet, for without that latter, the former could and would not exist. Yar and Steinmetz, 2019 proposed: It is the Internet that provides the crucial electronically generated environment in which cybercrime takes place. Cybercrime as previously defined, can incorporate the use of computers to support criminal acts around the world. Likewise, cybercrime is distinguished from computer crime, which is an umbrella term for the various crimes committed using the World Wide Web (WWW). According to David

Weissbrodt (2013), computers and the Internet are an improvement to workflow efficiency, but cybercrime is tied together through the vulnerability of computers and the Internet. In fact, computers were brought into society to help with business efficiency not to hinder agencies or even the overall general public.

As previously stated, an examination of cybercrime ought to begin with the Internet, for the simple reason that without the latter, the former could and would not exist. It is the Internet that provides the crucial electronically generated environment in which cybercrime takes place. Moreover, the Internet should not be viewed as simply as a piece of technology, more so as a kind of 'blank slate', that exits apart from the people that it does because people use it in particular ways and for a particular purpose (Snyder, 2001). Now, 'what' people do with the Internet, and 'how' they typically go about it, are crucial for understanding what kind of phenomenon the Internet is. Indeed, it is the kind of social uses to which we put the Internet that create the possibilities of criminal and deviant activity. To give one example, if people did not use the internet for shopping, then there would be no opportunities for credit card theft or crimes, that online activity exploit users would use to gather finical information. These opportunities can potentially put millions of people's sensitive data at risk like when the network of the Target retail chain, in the United States of America, was breached in 2013 and criminal actors absconded with potentially over a hundred million customers' credit and debit card information (Kassner, 2015). Similarly, it is because we use the Internet for electronic communication with friends and colleagues that the Love bug worm, which targeted email systems we use for that purpose, could cause billions of dollars in damage.

The Internet, as its name suggests, is in essence a computer network; or to be more precise, a 'network of networks' (Castells, 2002). A network links computer

together, enabling communication and information exchange between them. Many such networks of information and commination technology (ICT) have been in existence for decades- those of financial markets, the military, government departments, business organisations, universities and so on. The Internet provides the means to link up the many and divers' networks already in existence, creating from them a single network that enables communication between any and all 'nodes' (e.g. individual computers) within it.

Cyber became a 'plug-and-play' prefix that could be conjoined to any noun or verb to denote some relationship to computers and the Internet. Throughout the 1980s and 1990s uses of cyber proliferated include cybersoave, cybersex, cybershopping, cybersurfing, and, of course, cybercrime, among others (ironically, Norbert Wiener disapproved of cyber portmanteaus and said that they sounded 'like a streetcar making a turn on rusty nails' [as cited in Rid, 2016: 103]). Media theorist McKenize Wark (1997: 154) describes the problem as 'cyberhype' or the use of certain prefixes to give the illusion of explanatory power and significance to concepts. As she explains; the problem with cyber-hype is the easy assumption that the buzzwords of the present day are in some magical way instantly transformable into concepts that will explain the mysterious circumstances that generated the buzzwords in the first place. Cyber this, virtual that, information something, or the other. Viral adjectives, mutated verbs, digital nouns. Take away the number you first thought of and hey presto! Instant guide to the art of the new age, cutting edge, psychotropic anything-nut-postmodern what have you. Not so much a theory as a marketing plan. (Wark, 1997:154).

According to the Internet Crime Report 2016 of the FBI's Internet Crime Complaint Centre's (IC3), identity theft was ranked seventh with 16 878 victims, and

the loss of 58 917 398 USD (FBI, ICR, 2016) was recorded only in the USA. As per Internet Crime Report 2017, identity theft was the sixth biggest complaint with 17 636 victims and the loss of 66 815 298 USD only in the USA in 2017 (ICR, 2017). This clearly shows an increase in total number of victims and overall loss in dollars comparing it from the report of the previous year. The Federal Trade Commission's (FTC) annual summary of consumer complaints, for the year 2016, ranked identity theft third with a total of 399 225 complaints (FTC, 2017). In 2017, the FTC received a total of 2.7 million complaints of fraud, and identity theft was ranked second highest with 371 061 complaints (FTC, 2017). The identity theft became the top third complaint to the FTC in 2018, and the total number of identity theft complaints was 444 602 out of 3 million reports. Despite losing the ranking, the total number of identity theft complaints had increased by 11.3 per cent in 2018. Figure 1 shows the statistics from the Consumer Sentinel Network Data Book 2018 (FTC, 2019).

Soomro and Hussain (2019) suggested the above statical perception on identity theft or fraud. Again, below in Figure 1, shows the rate certain cybercrimes were reported from 2002-2018. Thus, giving the awareness that cybercrime itself is becoming a vast category of crime.

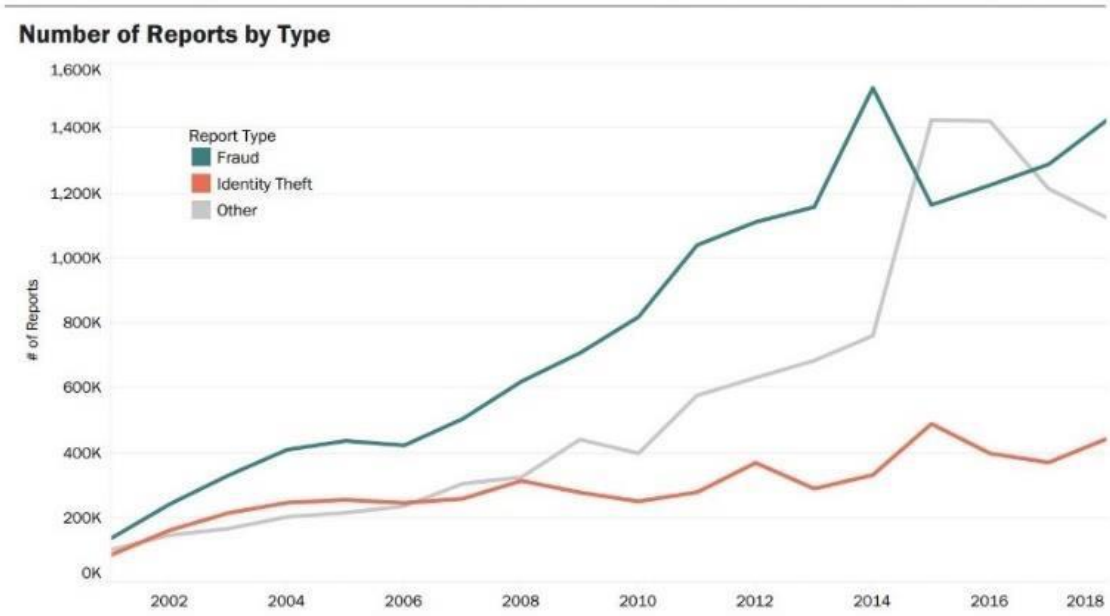


Figure 1. Identity theft complaints from 2002 to 2018 (Source: [21]).

However, at some point, collectively it was decided that not every online activity needs to be designated as ‘cyber’. Instead of ‘cybersurfing’, for instance, we now say’ browsing the web’, and ‘cybershopping’ is just simply ‘online shopping’. Even terms such like: ‘cyberspace’ and ‘cybersex’ have mostly fallen by the wayside. Cyber endures in certain forms, however, most notably in its application to harmful or illicit activities like cybercrime, cyberbullying, cyberharassment, and cyberterrorism. While originally the uses of cyber were ‘pregnant with promise of technology’, they have since come to connote the dangers of the Internet (Steinmetz and Nobles, 2018:3; see also Wall, 2012:5; Yar, 2014). Thus, the term cyber no longer seems to refer to the field of cybernetics but describes anti-social Internet activity. The negative connotations persist as politicians and pundits’ rail against the threats (both real and imagined) that ‘cyberattacks’, ‘cybercriminals’, and ‘cyberterrorist’ pose to our collective interests.

A search of the LexisNexis database indicates that news media interest in coverage of cybercrime has increased significantly over time, with 462 articles in 2000

and 4, 4640 in 2017. Evidence also suggests that news coverage of certain cybercrime topics has only increased over time. For instance, in their study of international news media coverage from 2008 to 2013, Jarvis et al. (2015:70) found that the number of news items discussing cyberterrorism had amplified over time, with a notable uptick after 2010. In his analysis coverage, Levi (2017,228: 373) notes that cybercrime 'is used as titillating entertainment which generates fear at the power of technology beyond the control of respectable society'. Thomas and Loader (2008:8) suggested that social transformation wrought by Internet technologies 'makes the future appear insecure and unpredictable', yielding a public and political overreaction.

Mainstream moral panics, which are fuelled by the media, lead to an excessive and unjustified belief that a particular individuals, groups or events present an urgent threat to society (Critcher, 2003). Yar (2012a) suggests that representations of the Internet in the popular imagination have increasingly come to by a 'cyber-dystopian' outlook, on that portrays the social effects of new technologies in overwhelming negative terms. For example, Internet-related instance of panics includes those over the effects of pornography in the mid-1990s, and more recently threats to child safety from paedophiles (Cassell and Cramer, 2008; Littlewood, 2003). The proliferation of such anxieties is perhaps best of which we currently find ourselves. This is not to suggest, however, that the dangers posed by cybercrime can simply be dismissed wholly unfounded. Nor is it to suggest that such widespread reactions ought to be simply ignored by criminologists.

Media representations, both factual and fictional, constitute an important criminological research topic in their own right; their careful examination enables us to uncover how the problem of cybercrime is being construed and defines, and how this shapes social and political responses to it (Butkvoic, et al., 2018), (Taylor, 2000; Vegh,

2002). Yet the weight of such representation can also serve to obscure the realities of criminal activity and its impacts, hindering rather than facilitating a balanced understanding of cybercrime. In addition, the difficulty that is exacerbated by the fact that cybercrime refers not so much to a single, distinctive kind of criminal activity, but more to a diverse 'range' of illegal and illicit activities that do share in common the unique electronic environment ('cyberspace') in which they take place. However, many considerable literature addresses or focuses on different kinds of cybercriminal activity, such as piracy, hacking, e-fraud, cyberbullying and cyberterrorism; each is examined and analysed considering the social, political, economic, and cultural context in which it takes shape.

The promise technological advances bear for teaching and learning is vast and largely unexplored. The breakthrough related to the big data paradigm and related advances in data mining, analysis, neural networks, and fuzzy logic techniques, suggest that it is feasible to establish more personalized and customized teaching and learning environments conducive to enhancing students' performance (Donalds and Osei-Bryson, 2018). With the development of mobile Internet and the micro-chip industry, smart terminals become an indispensable part of our daily lives, more than just a communication apparatus (Yang et al., 2013). More and more people prefer to use portable devices (e.g., smart mobile phone, or smart pad) to capture the scene they saw and the sound they heard instead of professional devices such as SLR (Single Lens Reflex) camera and audio recorder. However, the availability of numerous digital recording devices and a massive amount of recording data has brought new issues and challenges concerning the multimedia security (Jin et al., 2016). As a technique used for inspecting the originality, authenticity and integrity of

the digital data, multimedia forensics has become an attractive topic in information security (Stamm et al., 2013).

Furthermore, as previously stated, computers were introduced to act as a ‘helping hand’; however, with the rapid growth and ever-changing dynamics of digital technology computers have become just the opposite. With the effortless importance computers and the Internet bring to the work force and everyday lives, they also present countless additional opportunities for cybercrime. The understanding of cybercrime is simultaneously informed and obscured by political and media discussions of the problem. On the other hand, it is clear that the rapid growth of digital technology and the Internet has created unprecedented new opportunities for offending (Yar and Steinmetz, 2019). In addition, with the progression of digital technology, these opportunities have become abundant, and this growth aids the individuals who would cause an increase in cybercrime occurrences. Also, as mentioned before, it is important to remember that cybercrime can be any form of criminal activity utilising technology, whether it is a computer, smart phone, touch pad (tablet), or PDA (personal digital assistant or palmtop computer) (Marcum, 2019).

Lastly, the developments present serious challenges for law and criminal justice, as it struggles to adapt to crimes that no longer take place in the terrestrial world but in the virtual environment of cyberspace, which spans the globe through the Internet’s instantaneous communication, and affords offenders many new possibilities for anonymity, deception, and disguise (Yar and Steinmetz, 2019). The proposed data analytics research, which will be coded in this study, will further explore how latest technology is being compromised and greatly used to aid cyberstalking.

2.2. Cybersecurity

Cybersecurity is the only defence in one of the longest wars the world has ever known. Battles are fought daily against nation, states, organized crime, thieves, terrorists, and bored, but smart, kids. This war escalates every day as the battlefield grows. Code that can be exploited is everywhere, in watches and phones, smart bulbs and smart switches, thermostats and nuclear reactors. The cybersecurity defence that has been mounted is staggering. Ted Coombs, (2018) suggests: According to CSO from IDG, the expected cybersecurity budget for 2021 will be $6 trillion. It is not possible to create a perfect defensive barrier against everyone who might want to access computer systems that don't belong to them. The human element in computing has proven to be one of the weakest links in creating a defence.

Cybersecurity is multifaceted in the same way you might provide security for your own home, lights, an alarm, video surveillance, and locks. Protecting an enterprise is far more complex (Coombs, 2018). The job of protecting enterprises has changed over time to meet the demands of a changing threat landscape. With each advancement in cybersecurity, an equal or greater advance is made by those attempting malicious access (Coombs, 2018). Security professionals have realized that a bad actor with enough sophistication, time, and money wanting to breach their networks will ultimately succeed. In protecting a network and its endpoints, it's what you do next after an intrusion and how quickly you do it that matters.

The term cybersecurity is broadly used, whose definitions are highly variable, often subjective, and at times, uninformative. The absence of a concise, broadly acceptable definition that captures the multidimensionality of cybersecurity obstructs technology and scientific developments by reinforcing the predominantly technical

view of cybersecurity while separating regulations that should be acting in concert to resolve complex cybersecurity challenges. "Cybersecurity is the organization and collection of resources, processes, and structures used to protect cyberspace and cyberspace-enabled systems from occurrences that misalign de jure from de facto property rights" (Craigen, Diakun-Thibault, and Purse, 2014).

The 44th President of the United States of America, Barack Obama declared that “cybersecurity risks pose some of the most serious economic and national security challenges of the 21st century,” a position that has been repeated by leaders on countries from Britain to China (Singer and Friedman, 2014). Cybersecurity has become a matter of global interest and importance, with already more than 50 nations having officially published some form of strategy document outlining their official stance on cyberspace, cybercrime, and/or cybersecurity (Klimburgh, 2012). On September 23rd, 1982, Representative Don Edwards, a long-time member of the United States House of Representatives, presided over a congressional hearing to consider a new type of crime- “computer-related crime”. It is important to remember the United States of America is a large country in comparison to the United Kingdom. Therefore, how the US handles computer related crime may differ from the United Kingdom which can possibly be foreseen based off population or even the targeted value from criminals.

Moreover, looking at how the United Kingdom tackle cybercrime: from the National Crime Agency (NCA), the United Kingdom states cybercrime is a global threat. The national crime agency continues to state that criminals and the technical infrastructure (like the Internet) they use are often based overseas, making international collaboration essential. It is necessary to understand that cyberattacks are financially devastating and disrupting and upsetting to people and businesses.

The measure and complication of cyberattacks is wide ranging. Numerous technical tools mean that less technically capable criminals are now able to commit cybercrime and do so as awareness of the potential profits becomes more widespread. The evolving technical proficiency of malware means evolving harm as well as facilitating new crimes The NCA focus is on critical cyber incidents as well as longer-term activity against the criminals and the services on which they depend. In addition, the NCA also work closely with UK police, regional organised crime units, and partners in international law enforcement such as Europol, the FBI, and the US Secret Service to share intelligence and coordinate action.

Furthermore, Edwards set the scene: As the use of computers expands in our society, the opportunity to use computers to engage in or assist in criminal activities also expands. In response to this perceived problem, and several States has enacted legislation specifically aimed at computer fraud. The Federal Bureau of Investigation is attempting to enhance the security its computer facilities (Ellis and Mohan, 2019).

Edwards' above statement would be used through the next three-and-a-half decades, however, with slight tweaking here and there, using new terminology as the decades change. Recurrently, various representatives sitting in subcommittee meetings, policy forums, and other public venues would note that computers were progressively becoming ubiquitous and that their diffusion was leading to new types of harm that would ultimately call for new types of solutions. Although, at the start or at this time, the claimed harm(s) were speculative or theoretical; equally often, the calls of solutions followed publicized incidents that increasingly resonated in the public consciousness.

"Rarely has something been so important and so talked about with less and less clarity and less apparent understanding…I have sat in *very* small group meetings

in Washington… unable (along with my colleagues) to decide on a course of action because we lacked a clear picture of the long term legal and policy implications of *any* decision we might make." This is how General Michael Hayden, former Director of the CIA, described the cybersecurity knowledge gap and dangers it bestows (Singer and Friedman, 2014). It could be seen as the major part on this disengage is the consequences of the early experiences with computers, or rather the lack of them among too many leaders. For example, the youth of today are "digital natives," having grown up in a world where computers and the Internet have always existed and seem a very natural feature. However, the world is still mostly led by "digital immigrants," older generations for whom computers and the many issues the Internet age presents remain unnatural and often confusing (Singer and Friedman, 2014).

The Whitehouse (2011) outlined a cyber strategy that provides the stance of the United States of America (USA), on cyber-related issues and outlines a unified approach to the USA's engagement with other countries on cyber issues. The United Kingdom (UK) lists cybersecurity as a top priority and has committed £650 million over four years for a transformative National Cyber Security Programme (Minster for the Cabinet Office and Paymaster General, 2011). However, very few of these sources seem to make a distinction between the concepts of cybersecurity and information security or the relationship between them. The International Telecommunications Union (ITU) defines cybersecurity as follows:

Cybersecurity is the collection of tools, policies, security concepts, security safeguards, guidelines, risk management approaches, actions, training, best practices, assurance, and technologies that can be used to protect the cyberenvironment and organization and user's assets. Organization and user's assets include connected computing devices, personnel, infrastructure, applications,

services, telecommunications systems, and the totality of transmitted and/or stored information in the cyber environment. Cybersecurity strives to ensure the attainment and maintenance of the security of the organization and user's assets against relevant security risks in the cyber environment. The general security objectives compromise the following:

- Availability
- Integrity, which may include authenticity and non-repudiation
- Confidentiality ITU, 2008.

These definitions are very similar to that of information security. The international standard, ISO/IEC 27002 (2005), defines information security as: the preservation of the confidentiality, integrity and availability of information (ISO/IEC 27002, 2005, p.1). In the context of ISO/IEC 27002 (2005), information can take on many forms. It can be printed or written on paper, stored electronically, transmitted by post or electronic means, shown on films, conveyed in conversation, and so forth (ISO.IEC 27002, 2005, p.1). As well as Whitman and Mattord (2009) define information security as "the protection of information and its critical elements, including the system and hardwire that use, store, and transmitted that information" (Whitman and Mattord, 2009, p.8).

Cybersecurity, which serves to protect computer systems and data from malicious and accidental abuse and changes, both supports and challenges the reproducibility of computational science (Deelman, Taufer, Stodden, and Welch, 2019). Again, as previously mentioned cybersecurity is a broadly used term, whose definitions are highly variable, often subjective, and at time, uninformative. The Merriam-Webster defines cybersecurity as the: "measures taken to protect a computer or computer system (as on the Internet) against unauthorized access or attack" (Perkins and Weiss, 2013). Additionally, as cited earlier, Craigen, Diakun-Thibault,

and Purse, 2014 introduced a new definition: “cybersecurity is the organization and collection of resources, processes, and the structures used to protect cyberspace and cyberspace-enabled systems from occurrences that misalign de jure from de facto property rights.” However, the term cybersecurity like suggested before, has been the subject of academic and popular literature that has largely viewed the topic from a particular perspective.

The problem with networks, clouds, computers, and connected devices is that they are relevant to the public’s daily lives. It would be much easier to protect all these things if you didn’t have to introduce people into the equation. Educating the people that use your network against those trying to trick their way in could go a long way to improving security. Social engineering is a confidence scam that convinces unsuspecting people to provide information to bad people trying to steal their information. Social engineering is usually the first step in an otherwise complex attack. Think of it as opening the door to someone you believe is a trusted friend. You can easily underestimate the impact of social engineering on data security because it doesn’t feel technical enough or software- based. Between 66 and 84 percent of all network intrusion contains a social engineering factor (Coombs, 2018).

2.2.1 Phishing or Hacks

The number of large releases of personal information through hacks over the last few years means that almost everyone who has ever used a computer has had some aspect of their personal information stolen. Quite simply, bad people know something about you. Add to this, the fact that “privacy is dead” due to the amount of legal information collected by marketers, social networks, and all the other ways information is collected, sold, and shared about everyone. A big issue is that social

networks have made sharing your personal information an art. We love sharing our pictures of pets, where we eat, where we travel, and the names of all our friends and family. How many people use their dog's name in a password? It's just ripe with information. It's not a big task for a well-financed team to correlate all that data with the end-result being that malicious people know enough about you now to be experts at conning you.

Phishing is the further attempt to collect sensitive information about you that may not be readily available online. There is an endless list of phishing examples, while phishing has been around for a long time, a more targeted attack based on research specific to the victim called '*spear phishing'* has become popular in recent times (Coombs, 2018). Email, social media messages, or other information can be sent to you, and you believe the messages are from a trusted source such as an employer, frequently visited website (also known as a watering hole attack), a friend, person you know, or from companies where you regularly do business (Coombs, 2018). The information used to create these kinds of attacks come from stolen information. As previously stated, there are many forms of phishing.

Computers today are being misused for illegal activities like MasterCard fraud, spams, and so on, which invade our privacy and offend our senses. Criminal activities within the cyberspace are on the increase. The most dangerous frauds that cause in day-to-day banking activities are phishing, a criminal activity using social engineering techniques (Sharma, 2020). Phishers tend to target to fraudulently acquires user data information such as credit card details, passwords, and so on, by fraudulently representing themselves as a trusted entity. Communications connoting to be from popular social Internet sites, online payment processes or IT administrators are

commonly accustomed to lure the unsuspecting public (Sharma, 2020). Phishing links may contain websites that are infected with malware.

Sharma, (2020) introduces a list of types of Cyber Phishing:

1. Hand over Sensitive Information: the aim behind sending these messages is to acquire the important data of the user such as username, password, etc, to breach a system or account.
2. Download Malware: phishers sort of a lot of spam, these sorts of phishing emails aim to urge the victim to infect their own computer with malware.
3. Spear Phishing: when the phishers aim to draft a message to appeal a specific individual. It's called spear phishing. Phishers identify their targets and use spoofed addresses to send emails which anyone could believe like they're coming from co-workers or a trusted person and get tricked.
4. Whaling: whale phishing is a form of spear phishing aiming at the big fishes such as CEOs or other high-value targets. Generally, these scams target company board members, who are considered particularly accessible.

The larger your enterprise, the higher the chance that someone in the organization will fall victim to social engineering. The answer to protect against social engineering is education and awareness. Teach employees to look at the URL of origin, search websites about common scams, or even call the organization on the phone to verify the request before providing information (Coombs, 2018). In addition, the term cybersecurity is a revealing term to be explained for this research. Previously mentioned, however, it is imperative to understand that cybersecurity is used or involved in the same way people might stipulate security in their own home. Again, Coombs suggests that certain securities for the home might include: lights, an alarm,

possible video surveillance, and locks. Nevertheless, securing businesses and/or their reserved information is far more intricate (Coombs, 2018).

Craigen, et al., 2014: stated there is spectrum of technical solutions that support cybersecurity. However, these solutions alone do not solve the problem; there are numerous examples and considerable scholarly work that demonstrate the challenges related to organisational, economic, social, political, and other human dimensions that are inextricably tied to cybersecurity efforts (e.g., Goodall et al., 2009; Buckland et al., 2010; Deibert, 2012). For instants, how the threat landscape changes over time with the prompt evolution of new digital technology, thus protecting that same enterprise becomes rather difficult. Alongside that being stated, it is imperative to remember; each or all the progressions that are made in cybersecurity.

2.3. Cyberstalking

Stalking behaviours performed against former and current intimate partners account for many reported stalking situations, are continuously increasing, and can result in physical, psychological, and financial distress (Smoker and March 2017). The rise of technology has led to increased access to personal information and thus has facilitated the ease of stalking an intimate partner online (i.e., cyberstalking). However, the literature indicates a lack of clarity regarding predictive factors of perpetration of intimate partner cyberstalking behaviour (Smoker and March 2017).

Cyberstalking is the repeated unwanted relational pursuit of an individual through communication technologies, such as computers, tablets, and smart phones (Goodno, 2007; Reyns et al., 2012). Internet technologies are enticing platforms for stalkers because they create unique opportunities for perpetration (Nobles et al., 2014; Reyns, Henson, & Fisher, 2011). Although, cyberstalking can have numerous

definitions. Gibson (2019), defined cyberstalking as follows: stalking via some form of electronic medium such as email (Finn, 2004; Fox, Nobles, & Fisher, 2016; Strawhun, Adams, & Huss, 2013), or social media platforms such as Facebook and Twitter (Bennett, Guran, Ramos, & Margolin, 2011; Fox et al., 2016; Henson, Reyns, & Fisher, 2013; Marcum Higgins, & Ricketts, 2014; Nobles, Reyns, Fox, & Fisher, 2014; Reyns, Fisher, & Randa, 2018; Strawhun et al., 2013). In addition, as previously mentioned cyberstalking is the stalking of another through methods of electronic access and communication, such as, with the use of hidden webcams, GPS devices, and Spyware to monitor victim's behaviour, and pursuit and contact under anonymity through fake online pro-files (Sheridan & Grant, 2007; Shorey, Cornelius, & Strauss, 2015).

Stalking has been well recognised in the academic and practitioner literature; however, with the advent of technologies such as social media, a new threat has emerged, cyberstalking. An increased reliance of individuals on interpersonal contact in cyberspace has resulted in a corresponding increase in possibility of cyber-based personal intrusion, referred to as cyberstalking (McFarlane and Bocij 2005). This interpersonal intrusion in cyberspace includes behaviours or repeated events such as: repetition of an unwanted act of attention, invasion of personal privacy, as well as evidence of threat and/or fear (Spitzberg and Hoobler 2002). Additionally, the US Government considers cyberstalking to be ‘‘the use of the Internet, email, or other electronic communications devices to stalk another person’’ (US Attorney General 1999). For this reason, it can be understood that cyberstalking is a digital form of stalking and as such presents several commonalities in defining what constitutes an act of cyberstalking, yet it is quite distinct from traditional stalking behavior (Goodno 2007; Reyns et al. 2011; Spitzberg and Hoobler 2002).

Stalking is often a form of repeated behaviour (e.g. abuse is posted daily), which is unwanted by the victim and is abusive, threatening and causes a sense of danger and strong fear. This can include efforts to make contact directly “face to face” or indirectly on the Internet (Tomaszek 2012, p. 138). Dhillon and Smith (2019) suggested: stalking has been well recognised in the academic and practitioner literature; however, with the advent of technologies such as social media, a new threat of cyberstalking has emerged.

Stalking itself can be generally defined as stated above, however traditional stalking behaviour is contextualised by four following characteristics as suggested by Spitzberg and Hoobler (2002):

- First, it must be a repeated event, not a singular incident.
- Second, a person’s relative right to reasonable personal privacy must be violated.
- Third, evidence of threat must exist.
- Lastly, the threat is not solely limited to the person, but can expand to their property or social network.

These four characteristics signify a useful basis for distinguishing between a stalking and non-stalking-related acts; however, it must be implicit that cyberstalking adds an additional layer of complexity. According to Goodno (2007), there are five keyways in which cyberstalking differs from traditional stalking, representing a unique threat that must be addressed.

- First, those engaging in cyberstalking can use the Internet to instantly harass victims with much broader effect.

- Second, cyberstalking does not require a physical component; hence, no physical presence is required to commit the act.
- Third. Cyberstalkers usually remain completely anonymous.
- Fourth, as previously mentioned, due to reasons two and three, cyberstalkers more easily impersonate their victims.
- Lastly, cyberstalkers can encourage third-party bystander harassment, inciting others to commit the harassment in their place (Goodno 2007).

Stalking inside the cyber world by using the social media or any other online medium, which may cause feelings of irritation, abuse, and emotional anxiety to the victim, can be described as cyberstalking (NW3C, 2015). Whilst there is no strict legal definition of stalking (CPS, 2014), the act of stalking involves “harassing or persecuting (someone) with unwanted and obsessive attention”. The UK legislation as of 2012, has added acts of stalking and cyberstalking to the Protection from Harassment Act 1997, and recognises stalking should become recognised as a criminal offence. As previously mentioned, stalking has been well recognised in the academic and practitioner literature; however, with the advent of technologies such as social media, a new threat has emerged, cyberstalking. An increased reliance of individuals on interpersonal contact in cyberspace has resulted in a corresponding increase in possibility of cyber-based personal intrusion, referred to as cyberstalking (McFarlane and Bocij 2005). This interpersonal intrusion in cyberspace includes behaviours such as repeated events (repetition of an unwanted act of attention), invasion of personal privacy, as well as evidence of threat and/or fear (Spitzberg and Hoobler 2002). Additionally, the US Government considers cyberstalking to be ‘‘the use of the Internet, email, or other electronic communications devices to stalk another person’’ (US Attorney General 1999). For this reason, it can be understood that cyberstalking is a

digital form of stalking and as such presents several commonalities in defining what constitutes an act of cyberstalking, yet it is quite distinct from traditional stalking behaviour (Goodno 2007; Reyns et al. 2011; Spitzberg and Hoobler 2002).

Further the role of technology in digital stalking offences known as cyberstalking, cyberstalking has also been highlighted within legislation (Horsman and Conniss, 2015). However, the prosecution of such cyberstalking offences is reliant on the forensic analysis of devices capable of communication with a victim; and with the proliferation of anonymous communication services, it is becoming extremely difficult for digital forensics specialists to analyse and detect the origin of the cyberstalker's messages (Horsman and Conniss, 2015). Thus, the importance of data mining and data analytics on cyberstalking would emphasise and/or help with the detection of origin of the cyberstalker's messages on either a membership only or public platform.

Moore (2018) suggested, in the USA alone that one woman out of twelve and one man out of forty-five could be stalked in their lifetime. Moore's article suggested that females between the ages of 18 and 29 were mainly the victims of cyberstalking. Conversely, research from a survey at the University of Pennsylvania shows that 56 per cent of cyberstalking victims were, however, male. According to the Bureau of Justice Statistics (Bureau of Justice Statistics: BJS), every 14 out of 1000 persons at the age of 18 were victims of stalking and around 1 out of 4 victims' complaints were about some sort of cyberstalking in the form of e-mail and instant messaging. Lastly, Duggan (2017) illustrated, that one out of ten Americans had experience(s) of online harassment and 7 per cent of American adults had faced a form of cyberstalking.

"As social networking becomes more a part of our daily lives, individuals find this technology an attractive vehicle to perpetrate cybercrimes" (Ackerman & Schutte,

2015, p. 2). This could be due to the power of anonymity for some but as shown with Twitter harassment, abuse and threats can occur as well. Furthermore, social media networking services such as Twitter and Facebook afford access to a larger volume of potential victims (Sen, 2013). Stalkers often demonstrate a sustained obsession with their victim that can be stimulated by the ease of access to, and the volume of personal information placed online (Casey, 2011). Confusion surrounds the term cyberstalking, because it is used aside colloquial phrases such as Facebook stalking or friend stalking. In public discourse, Facebook or friend stalking refers to surreptitious online information seeking behaviours (Lyndon, Bonds-Raacke, & Cratty, 2011; Parsons-Pollard & Moriarty, 2009; Tokunaga, 2011, 2016), whereas cyberstalking involves the repeated pursuit of a targeted individual over the Internet (Reyns et al., 2011). Cyberstalking is sometimes viewed as an analogy to offline stalking but enacted through the Internet (Tjaden, 2014).

Although many types of cyberstalking have been documented, behaviours involving the pursuit of unwanted relationships using technologies are of main interest to those who study interpersonal violence. McFarlane and Bocij (2003) used the term “intimate cyberstalkers” to describe the group of individuals, which includes ex-intimates and infatuates, who use technology for unwanted relational pursuit. Intimate cyberstalking can involve psychological intimidation and threats made between people in existing relationships, but individuals not involved in a relationship can also be pursued (Southworth, Finn, Dawson, Fraser, & Tucker, 2007). For example, a motive as to why cyberstalking is being used is: the individual’s appearance, facial expression, body language, voice, and dress or demeanour, is not shown during Internet transactions (Smith and Urbas, 2001, 1). This finding illustrates or mirrors face to face stalking, but the growth of digital technology compelled or elevated stalking

into a more invasive crime conveyed using the Internet. This criminal act is potentially growing widely due to the connection and the use of digital technology and social media networks; however, the questions remain of how cyberstalking is being detected or flagged and how are potential victims being protected?

Cyberstalkers tend to use the same behaviours or patterns as stalkers, making cyberstalking just as fearful if not more so for the victim, because of the correlated use of digital technology. Cyberstalking is a significant challenge in the era of the Internet and technology (Dhillon and Smith, 2019), and institutions, governments, and social media platforms struggle in how to manage it and where to allocate resources. This is the reason why it is imperative to recognise how to prevent the problem of cyberstalking and how it can be countered with security measures. Lastly, recent literature and research has shown that offences linked to cyberstalking have focused on the comparison between offline stalking and cyberstalking itself; therefore, the necessity to increase understanding of the technological means of detecting and gathering evidence in cases of cyberstalking is paramount (Frommholz, al-Khateeb, Potthast, Ghasem, Shukla, and Short, 2016). Data analytics such as data mining in computer science is a promising technology that could be used to detect cyberstalking.

The knowledge gap that surrounds cyberstalking is assisted with previous research helps to develop a greater understanding of cyberstalking such as what legally constitutes cyberstalking, the role society plays in governing cyber-based misbehaviours, and regulatory issues governments and institutions face when attempting to prevent it. However, the literature does not comment on how actual measures can be developed to detect cyberstalking.

Lastly, cyberstalking thus entails the same general characteristics as traditional stalking, but in being transposed into the virtual environment it is fundamentally

transformed, via the Internet. The nature of this transformation is dependent upon what aspects of the Internet are exploited. The Internet is used by both those with an interest in efficient exercises of "traditional" criminality as well as those attuned to the possibilities of altogether new forms of criminality (Grabosky, 2000). In addition, Ogilvie (2002) suggests: 'in both instances, we need to understand what the Internet actually 'is' if we wish to determine the potential for criminality it entails. At present, attempts to protect the wider community from cybercrimes are hindered by a failure on the part of policy makers to appreciate that the Internet offers access to domains beyond the reach of traditional legislative frameworks.' This means that before understanding the nature of cyberstalking, we need to first understand the nature of the Internet. At a superficial level, the Internet is conceptually very simple. All that is happening is we have a means of efficiently transferring digitised data. Ogilvie (2000) explains that there are three major ways in which these data exchanges can be categorised. These three major ways in which data exchanges will be clarified further in the next section.

2.4. Social Media Analytics

2.4.1 Introduction

Soomro and Hussain, 2019 brought into consideration, social media-related cybercrimes, and techniques for their prevention. Soomro and Hussain (2019) mention: according to "Criminal Use of Social Media" white paper from the National White-Collar Crime Centre (NW3C, 2013), social media has been on rise in past several years, which changes the communicational landscape. Social media sites, such as Facebook, Twitter, Instagram, and YouTube, have millions of active users. Using these websites, people communicate instantaneously with each other with

convenience. Social media sites are used by people to communicate with each other, and by the public sector for advertisement and recruitment of new employees. Statista's data on social media users as of January 2019 are shown in Figure. 2.

In Figure 2, which is the figure below, shows social media users and the different social media networks and or platforms that are being used as of January 2019.

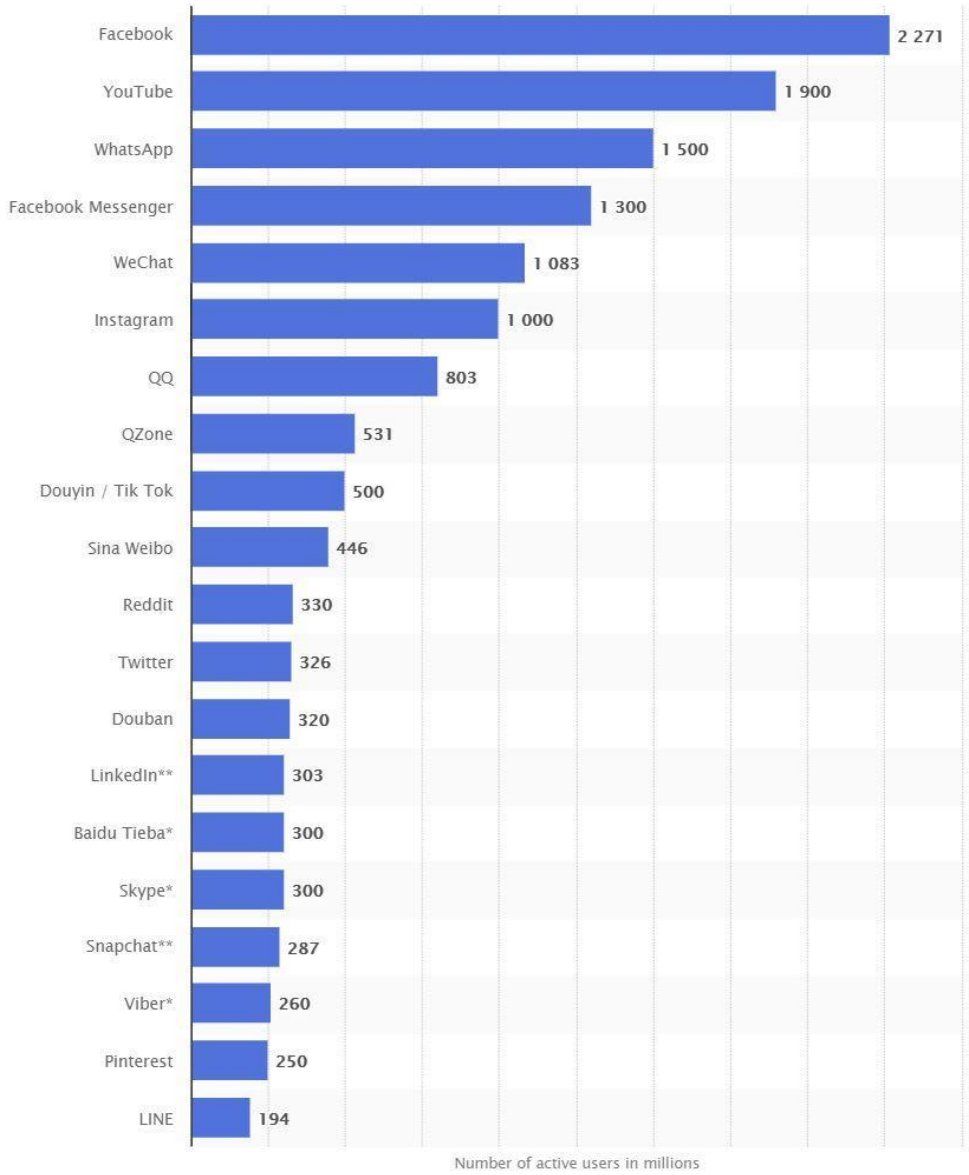


Figure 2. Social media users as of January 2019 (Source: [3]).

2.4.2 Social Media Categories

As previously mentioned, there are three major ways in which these data exchanges can be categorised. The three ways that categorises the Internet are as follows (Ogilive. 2000):

- Category One: The Internet as a medium of convenience Data may be transferred directly from sender to a nominated and “willing” recipient. Person to person email communications are the most obvious example of this type of data exchange.
- Category Two: The Internet as a Medium of Control. Data may be exchanged in an interaction involving an unwilling and/or unknowing party being manipulated by a usually unknown and effectively invisible external party. Exploitation of the “back door” into the Windows operating system in order to take over control of another computer connected to the Internet is the obvious example of this second type of data exchange.
- Category Three: The Internet as a medium of “range” enhancement Data may be electronically positioned in such a way that any number of data “seekers” may locate and obtain it. Web sites that are “hit” by “net-surfers” are the obvious example of this third type of data exchange. Other examples include Internet Relay Chat (IRC), Multiple-User Dungeons (MUDs) and Multiple Object-Oriented domains (MOOs).

Each of these three types of data exchange mechanisms constitutes a particular form of virtual social interaction that may or may not accompany an interaction ‘in real life’ (IRL). Some might argue that the distinction between the virtual world of cyberspace and the more substantial IRL world is false or misleading, that in both cyberstalking and IRL stalking, real people are communicating whether it be by computer connections, telephone, letters, face to face contacts, or whatever.

However, the extent to which (virtual) cyber interactions both resonate with (IRL) pre-Internet endeavours and are amenable to regulation, is far from uniform. The cyber/IRL distinction is therefore useful for clarifying the difference between cybercrimes that are extensions of traditional criminal behaviours, and cybercrimes that are wholly distinct from traditional forms of criminality.

For example, the use of the Internet to distribute illegal pornography, is a good example of a cybercrime that closely resembles the “real” world. Digitised images are transferred on a commercial basis between providers and purchasers. Essentially, all that is occurring is that the Internet is providing a convenient new forum for a particular aspect of the “hidden” or “black” economy (as well as the legitimate trade in legal images). Digital data are transferred instead of pictures in brown paper envelopes. However, there are activities such as cyberstalking, which owe something to the real world and something to the virtual world as well. In thinking about cyberstalking in terms of the three categories of cyber interaction we can obtain a sense of how the movement from IRL crime to virtual crime may entail a shift from the controllable to the uncontrollable. Cyberstalking provides an especially illustrative example of this shift because of the way it can occur as an instance of each of the three types of cyber interaction.

The internet and mobile technologies have been the primary force behind the rise of social media, providing technological platforms for information dissemination, content generation, and interactive communication. From an application perspective the ones relating to social media are the most popular. For instance, Wikipedia (collective knowledge generation), Facebook (social networking), YouTube (social networking and multimedia content sharing), Digg and Delicious (social browsing, news ranking, and bookmarking), Second Life (virtual reality), and lastly Twitter (social

networking and microblogging) to name a select few. Since social media is already a critical part of the information ecosystem and as social media platforms and applications gain widespread adoption with unprecedent reach to users such as: consumers, voters, businesses, governments, and non-profit organisations alike, interest in social media from all walks of life has been skyrocketing from both application and research perspectives. However, from a tool perspective, an array of Web-based applications defined the way social media functions. Examples, include weblogs, microblogs, online forums, wikis, podcast, life streams, social bookmarks, Web communities, social networking, and avatar-based virtual reality.

2.4.3 Social Media Analytics

Social media analytics "is concerned with developing and evaluating informatics tools and frameworks to collect, monitor, analyse, summarize, and visualize social media data … to facilitate conversations and interactions … to extract useful patterns and intelligence… (Fan and Gordon, 2010)". In the early days of social media, agencies would monitor customers' posts on a business's own website to try to identify and manage unhappy customers. With the explosion in the number of social media sites and volume of users on them, monitoring alone is not enough to render a complete picture of how a company is doing. Consider the pervasiveness of social media:

- Social networking is the most popular online activity.
- 91 per cent of adults online are regular users of social media; and
- Facebook, YouTube, and Twitter are the second, third, and eighth most trafficked sites on the Internet, as of April 2014 (Fan and Gordon, 2014).

Fan and Gordon (2014) continue with stats measured with percentages on social media analytics. However, it is important to remember not even these statistics do not fully account for the influence social media has on our lives. Users spend more than 20 per cent of their time online on social media sites. Facebook alone has a worldwide market penetration rate over 12 per cent of the entire online population; in North America it is 50 per cent. These rates are growing quickly, with Facebook alone gaining 170 million new users between the first quarter of 2011 and the first quarter of 2012, an increase of 25 per cent. Facebook mobile use is growing even more quickly, at a 67 per cent annual clip, as of summer 2013 (Fan and Gordon, 2014). The amount of information seen by all these users on a typical day gives a clearer indication of the enormous influence of social media. As of October 2012, Facebook's nearly one billion active users collectively were spending approximately 20,000 years online each day. In the same period, YouTube reported more than one billion views, or 500 years of video (spread among 800 million unique users), and 140 million active Twitter users sent more than 340 million tweets (Fan and Gordon, 2014).

Nonetheless, these are not simply passive uses. YouTube's analysis of its videos indicates 100 million people take some sort of "social action" each week, by, saying, liking, disliking, or commenting on what they see; these actions doubled from 2012 to 2013 (Fan and Gordon, 2014). Facebook integrates social actions in its online ads today by, for instance, allowing users to see if their friends have liked or voted on products being advertised. Likewise, hashtags on Twitter (as well as other social-media platforms) give users another quick and easy way to express their likes, dislikes, interests, and concerns, presenting further opportunities (or challenges) to businesses striving to use them (Fan and Gordon, 2014).

Fan and Gordon (2014) made a model for social media analytics. Social media analytics involves a three-stage process: “capture,” “understand,” and “present”. Below are they have defined each section:

- **Capture**, for a business using social media analytics, the capture stage helps identify conversations on social media platforms related to its activities and interests. This is done by collecting enormous amounts of relevant data across hundreds or thousands of social media sources using news feeds and APIs or through crawling. The capture phase covers popular platforms (such as Facebook, Foursquare, Google+, LinkedIn, Pinterest, Twitter, Tumblr, and YouTube), as well as smaller, more specialized sources (such as Internet forums, blogs, microblogs, wikis, news sites, picture-sharing sites, podcasts, and social-bookmarking sites). An enormous amount of data is archived and available to meet businesses’ needs. To prepare a dataset for the understand stage, various pre-processing steps may be performed, including data modelling, data and record linking from different sources, stemming, part-of-speech tagging, feature extraction, and other syntactic and semantic operations that support analysis. Information about businesses, users, and events, as well as user comments and feedback and other information, is also extracted for later analytical modelling and analysis. The capture stage must balance the need to find information from all quarters (inclusivity) with a focus on sources that are most relevant and authoritative (exclusivity) to assist in more refined understanding.
- **Understand**, when a business collects the conversations related to its products and operations, it must then assess their meaning and generate metrics useful for decision making the understand stage. Since the capture stage gathers data

from many users and sources, a sizeable portion may be noisy and thus have to be removed prior to meaningful analysis. Simple, rule-based text classifiers or more sophisticated classifiers trained on labelled data may be used for this cleaning function. Assessing meaning from the cleaned data can involve statistical methods and other techniques derived from text and data mining, natural language processing, machine translation, and network analysis. The understand stage provides information about user sentiment how customers feel about a business and its products and their behaviour, including the likelihood of, say, purchasing in response to an ad campaign. Many useful metrics and trends about users can be produced in this stage, covering their backgrounds, interests, concerns, and networks of relationships. Note the understand stage is the core of the entire social media analytics process. Its results will have a significant effect on the information and metrics in the present stage, thus the success of future decisions or actions a business might take. Depending on techniques used and information sought, certain analyses may be pre-processed offline while others are computed on the fly using data structures optimized for anticipated ad hoc uses. Analysts and business managers may participate directly in the understand stage when visual analytics allows them to see various types and representations of data at once or create visual “slices” that make patterns more apparent.

- **Present**, in this last stage, the results from different analytics are summarized, evaluated, and shown to users in an easy-to-understand format. Visualization techniques may be used to present useful information; one commonly used interface design is the visual dashboard, which aggregates and displays information from multiple sources. Sophisticated visual analytics go beyond the

simple display of information. By supporting customized views for different users, they help make sense of large amounts of information, including patterns that are more apparent to people than to machines. Data analysts and statisticians may add extra support (Fan and Gordon, 2014).

2.4.4 Social Media Networks or Digital Technology

Social media (SM) is a set of Internet-based applications that is grounded by the idea of Web 2.0 (E. Gilbert & Karahalios, 2009). SM was initially used around 2004 to describe contents and applications that can be continuously modified and altered by users in many ways through participation and collaboration, rather than traditionally created, prepared, and published by only individuals (Kaplan & Haenlein, 2010). The broad utilization of available software and hardware to access social media platforms over the Internet led to the creation and exchange of user-generated content. Ellison (2007) listed three aspects to define social media that they referred to as “social network site” as web-based services:

- First, individuals can create their public or semi-public profile.
- Second, these individuals can connect to others to form a network.
- Last, these individuals can view and relate to other users and their activities, which are publicized in their network.

The terms social media and social media sites have been used interchangeably. In this paper, the term social media refers to any social network sites that have all the three aspects as per Ellison (2007). The examples of social network sites that generate a large amount of un- structured data are Facebook, Twitter, Instagram, LinkedIn, blogs, wikis, and YouTube. Social media big data along with the progress in computational tools have emerged as the key to crucial insights into human behaviour and are continually stored and processed by corporations,

individuals, and governments (Manovich, 2011). The most common applications of big data for social media are trend discovery, social media analytics, sentiment analysis, and opinion mining. For instance, social media assists organizations to obtain customers' feedback regarding their products, which can be used to modify decisions and to obtain value out of their business (Katal, Wazid, & Goudar, 2013; Wu, Zhu, Wu, & Ding, 2014). Studies confirmed that most of the existing approaches to social media big data analysis rely on machine learning techniques (Cambria, Rajagopal, Olsher, & Das, 2013).

The popularity of the Internet and the advent of the Web 2.0 technologies have transformed the contents of the web from publisher- to user-created contents (Alexander, 2006). Such existence has assisted in publishing contents without the needs of programming. Today, interesting topics, reviews, and opinions from Web 2.0 and social media can easily be accessible globally via the Internet in real time by anyone. Moreover, the proliferation and adoption of social media have provided extensive opportunities and challenges for researchers and practitioners. More than a billion of people around the world are using social media platforms that generate overwhelming unstructured data in relatively short timescales. The huge amount of data generated by users is the result of the integration of their background details and daily activities in such platform. This massive amount of generated data referred to as “big data” has been intensively researched recently.

The big data from the huge amount of the dataset collected in either structured, semi-structured and/or unstructured format have been researched in various domains, such as healthcare, astronomy, social web, and geoscience (Hashem et al., 2015). Social media contents, such as tweets, comments, posts, and reviews, have

contributed to the creation of big data extensively from either platform providers or different websites (Kwon, Lee, & Shin, 2014; Lyu & Kim, 2016).

The use of social media and the Internet has become very mainstream in today's society. According to the authorities, cyber threats are one of the main threats to national security in the United Kingdom (UK) that has significant impact on the IT infrastructure. As the UK economy largely depends on its IT infrastructure to digitally support businesses, commerce and private citizens, industries and academics urgently seek security strategies and measures to combat cyber threats.

Online social media platforms have gradually gained ground, allowing individuals to share their thoughts and emotions around a variety of happenings in their daily lives. The emotion and language used in social media posts, conversations and messages contain vital information useful for understanding user features, such as, likes and dislikes, and post features. These features may indicate whether users have developed 'rapport' in short interactions and longer relationships, and when rapport is lost. For instance, many positive conversations between two individuals may be indicative of the development of close/good rapport among them.

Social media has granted the potential (cyber) stalker the access to obtain an individual's personal information. The number of large releases of personal information through hacking into IT systems over the last few years means that almost everyone who had ever used a computer has had some aspect of their person information stolen (Home Office, 2017). Social media has made sharing most of a person's personal information an art form; from sharing personal photos, to where people go for dinner, to where they travel, even to the names of their family members and friends, every second is of their lives is documented. It also includes, the amount

of legal information collected by marketers, social networks, and many other ways information is collected, sold, or shared through social networks. Therefore, if hostile people want to know something about a social media user and adding the notion that 'privacy' is dead, social media networks are assisting towards that person accessing their information as well as their everyday life via the Internet. Moreover, Soomro and Hussain, 2019 as previously mentioned and shed light on the factors of social media use with the astonishing numbers: according to [2], in March 2019 the number of Internet users reached 4 168 461 500, i.e., 50.08 per cent penetration of world population and according to [3], in 2019 there were 2.77 billion social media networking users worldwide., 35.9 per cent of global social media networking penetration and it is expected that in 2021 this number will reach 3.02 billion.

The internet age has brought with it a slew of tools and research, which allows potential stalkers, from ex-lovers, friends, acquaintances, to even complete strangers, to follow a person's life in much detail without their consent. Especially, still the majority of cyberstalking cases concern two (or more) ordinary people who were loves or acquaintances (Eterovic-Soric, Choo, Ashman, Mubarak, 2017). Stalking is a pattern of behaviour with many definitions: 'the legal definitions differ from country to country, even state to state, (which proves why data analytics can be useful and help prevent and detect potential offenders/stalkers') influenced by local stalking cases as laws were being enacted (similar to other cyber-related offences, such as online child exploitation' (Hillman, Hooper, Choo, 2014). It is common for friends and family to be abused by a potential stalker attempting to gain access to their primary target, thus causing them to be in social isolation.

'Studies conducted before the Internet became widespread reported that a woman was 12-14 per cent likely to be stalked over her lifetime, and a man was 4-7 per cent

likely to be stalked over his lifetime' (Sheridan, Blaauw, Davies., 2003;4(2):148-62). The internet age has brought about an unprecedented level of interconnectedness and the advances in communications technology have enabled friends and colleagues to keep in touch wherever they are in the world (Eterovic-Soric, et al, 2017). Digital technology and the use of social media networks include all types of electronic equipment and applications. For instance, a case in the United Stated called: Mary's Case. Her case is as follows: Mary was cyberstalked by an individual who she had no contact within her daily life. This individual watched her 'online life' and used digital technology to his advantage to cyberstalk Mary. The use of digital technology prolonged the fear of not knowing where her stalker was within the world. Moreover, one of the crucial differences between cybercrime and traditional crime is imposed. "Traditional crime typically occurs in one space and has an impact on one set of victims, whereas cybercrime can have a global impact" (United Kingdom 2010, 5). Cybercrime again compared to offline crime is more difficult to define than traditional offline crime. The reasons for this are due to a computer or that device can be the cause, target or organiser of the crime, and the crime can take place on the computer alone or in other offline locations (Gordon and Ford, 2006, 13). It is for the reasons and many more that cyberstalking can become vastly recognised and committed worldwide.

2.4.5 The Internet Age

As previously stated, the Internet age has brought with it a slew of tools and research which allows potential stalkers, either from ex-lovers, friends, acquaintances, to even complete strangers, to harass, threaten and abuse. Whilst stalking is a pattern of behaviour where: 'the legal definitions differ from country to country, even state to state (in the US), influenced by local stalking cases as laws were being enacted

(similar to other cyber-related offences, such as online child exploitation' (Hillman, Hooper, Choo, 2014). This is where data analytics can be useful aid to help prevent and detect potential offenders/stalkers, as it is based on data, recording the volume of posting and abuse and velocity and frequency.

To continue, the Internet age has brought about an unprecedented level of interconnectedness and the advances in communications technology have enabled friends and colleagues to keep in touch wherever they are in the world (Eterovic-Soric, et al, 2017). Online social media represent a fundamental shift of how information is being produced, transferred and consumed. User generated content in the form of blog posts, comments, and tweets establishes a connection between the producers and the consumers of information. Social Media provides a connection between our social networks, personal information channels and the mass media. Social Media data in the form of user-generated content on blogs, microblogs like Twitter, discussion forums, product review and multimedia sharing websites present many new opportunities and challenges to both producers and consumers of information. Although there is a vast quantity of data available, the consequent challenge is to be able to analyse the large volumes of user-generated content and often implicit links between users, and to gain meaningful insights (Leskovec, 2011).

Moreover, social network analysis is used to model social network dynamics and growth (using such features as network density and locations of new node attachments) that help monitor business activity. Social network analysis is the primary technique for identifying key influencers in viral marketing campaigns on Twitter and other social media platforms. It is also used to detect sub communities within a larger online community (such as discussion forums), allowing greater precision in tailoring products and marketing materials. It is also useful in predictive modelling, as in

marketing campaigns aimed at consumers assumed most likely to buy a particular product (Bonchi, Castillo, Gionis, and Jaimes, 2011).

Social media has evolved over the last decade to become an important driver for acquiring and spreading information in different domains, such as business (Beier & Wagner, 2016), entertainment (Shen, Hock Chuan, & Cheng, 2016), science (Chen & Zhang, 2016), crisis management (Hiltz, Diaz, & Mark, 2011; Stieglitz, Bunker, Mirbabaie, & Ehnis, 2017a) and politics (Stieglitz & Dang-Xuan, 2013). Social media platforms offer many possibilities of data formats, including textual data, pictures, videos, sounds, and geolocations. Generally, this data can be divided into unstructured data and structured data (Baars & Kemper, 2008). In social networks, the textual content is an example of unstructured data, while the friend/follower relationship is an example of structured data. The progression of social media usage opens innovative opportunities for analysing several aspects of, and patterns in communication. For example, social media data can be explored to increase perceptions into issues, trends, influential actors, and other kinds of information. Golder and Macy (2011) analysed Twitter data to study how people's mood changes with time of day, weekday, and season. In the field of Information Systems (IS), social media data is used to study questions such as the influence of network position on information diffusion (Susarla, Oh, & Tan, 2012). "Traditional crime typically occurs in one space and has an impact on one set of victims, whereas cybercrime can have a global impact" (United Kingdom 2010, 5). It is for the reasons and many more that cyberstalking can become immensely well-known and committed globally.

2.5. Use of Machine Learning

2.5.1 Introduction

Artificial intelligence (AI) today is a world-changing tool that enhances humankind's abilities in several areas (Ted Coombs, 2018). We face the terrifying reality of what malicious hackers have done to our privacy and the grip of terror they keep us in. Never knowing when they'll use their evil prowess to destroy our credit or shut down the local power plant, what we need now are smart machines that help security professionals in the fight against this evil reality.

Computer and network intrusions have shut down airports and hospitals, interrupted commerce, and held people and businesses ransom for their data. The more data that's stolen the more power evil doers must create exploits that trick you into the foolish behaviour of clicking malicious links. Cybersecurity is one of the greatest challenges of this generation. It seeks to protect our world's data, ideas, and processes; thwart criminal enterprises that prey on our businesses; and exploit people around the world (Coombs, 2018). This area is one of the most understaffed industries in the world with unfilled cybersecurity positions to be about 1.5 million by the year 2020 (Coombs, 2018). Help is needed to make the current security professionals more efficient and augment their intelligence. This assistance is coming from AI (Coombs, 2018).

One of the things smart machines are good at is analysing data, such as text and images, by using a process known as *pattern recognition,* considered a branch of machine learning. Pattern recognition uses both supervised (training data) and unsupervised (no training data) to find patterns in data, either visual or textual. Bothe these branches of machine learning will be defined below. In addition, most visual pattern recognition is done using supervised learning algorithms. A significant number of training images are provided for the computer to learn and be able to recognize a pattern. Pattern recognition in text data is sometimes called *data mining.* One prime

example of this, that is used today is the Gmail auto response system that makes suggestions of email responses based on the content in your email.

Machine learning algorithms are programmes that can learn from data and improve from experience, without human intervention. Learning tasks may include learning the function that maps the input to the output, learning the hidden structure in unlabelled data; or 'instance-based learning', where a class label is produced for a new instance by comparing the new instance (row) to instances from the training data, which were stored in memory. "Instance-based learning' does not create an abstraction from specific instances.

However, Ted Coombs (2018) continues to bring forth information on algorithms to explain insights into machine learning. There is no single way to design machines that learn. The underlying code contains *learning algorithms* that are programs that extrapolate insights (intelligence) based on data provided to the computer. There are two basic categories of learning algorithms, supervised and unsupervised (Coombs, 2018).

2.5.2 Supervised and Unsupervised Learning

Supervised and Unsupervised Learning, defined by Ted Coombs (2018):

- **Supervised learning**: supervised learning is exactly what it sounds like. Someone supervises the input of information upon which the learning algorithm will arrive at a conclusion. Think of this like giving the computer a tutor. One of the most basic supervised learning algorithms is designed around a decision tree. This is the foundation of the expert system, a series of yes and no questions sufficient for the computer to arrive at some probable answer. With an expert system, a conclusion is derived based on the programmed inputs of

field experts. For example, diagnosing starter problems in a car will require the user to answer questions about the symptoms experienced when trying to start the car. Do you hear a click when you turn the key? Yes or No. Based on that answer, new questions along the tree are asked until the computer suggests, “Your battery is likely dead.”

- **Unsupervised learning**: unsupervised learning allows for the training of AI, using data that’s unlabelled and unclassified with the use of special algorithms that allow the AI to learn on its own rather than being spoon fed the data by a human. Two common unsupervised algorithms include the *apriori* and the *k-means* (Coombs, 2018).

Unsupervised learning models are used more when there is only an input for (X) and no corresponding output variables. They use unlabelled data to model the underlying structure of the data. James Le (web, 2019) states: three types of unsupervised learning.

- **Association:** is used to discover the probability of the co-occurrence of items in a collection. It is extensively used in market-basket analysis. For example, an association model might be used to discover that if a customer purchases bread, s/he is 80 per cent likely to also purchase eggs as well.
- **Clustering:** is used to group samples such that objects within the same cluster are more similar to each other than the objects from another cluster.
- **Dimensionality Reduction:** is used to reduce the number of variables of a data set while ensuring that important information is still conveyed. Dimensionality Reduction can be done using Feature Extraction methods and Feature Selection methods. Feature Selection selects a subset of the original

variables. Feature Extraction performs data transformation from a high-dimensional space to a low-dimensional space. Example: PCA algorithm is a Feature Extraction approach.

The usage of supervised learning is as follows, labelled training data to learn the mapping functions that turns input variables (X) into the output variable (Y). Nonetheless, it solves for f in the following equation: Y= f (X): this allows to accurately generate outputs when given new inputs. There are two types of supervised learning, they are classification and regression. Again, mentioned by, James Le (web, 2019):

- **Classification:** is used to predict the outcome of a given sample when the output variable is in the form of categories. A classification model might look at the unput data and try to predict labels.
- **Regression:** is used to predict the outcome of a given sample when the output variables is in the form of real values. For example, a regression model might process input data to predict the amount of rainfall, the height of a person, etc.
- **Ensembling:** is another type of supervised learning. It means combining the predictions of multiple machine learning models that re individually weak to produce a more accurate prediction on a new sample.

More advanced types of learning come from the use of training data. Unlike an expert system where specific answers are provided by experts, allowing a computer to learn by training provides unique capabilities. Feed the computer hundreds of thousands of cat photos and eventually it will be able to recognize a cat in a photo. This is a step up in learning because, it’s based on probability. Based on every other cat picture the computer has seen, it forms a probable idea of cats in a photograph.

Two common types of learning algorithms of this type are logistic regression and a back propagation neural network.

However, artificial intelligence (AI) applied to cybersecurity provides security professionals with an augmented ability to protect endpoints, data, and networks. By using sophisticated abilities to predict problems based on prior solutions and an ability to use natural language processing (NLP) to analyse unstructured data, unique solutions and detailed insight are provided to the Security Operations Centre (SOC) to quickly and cost- effectively stop intrusions or even prevent them before they happen (Coombs, 2018). It's also the only way to protect a network against malicious attackers also using artificial intelligence (AI).

Cybersecurity professionals use analytics to detect anomalies in network patterns, network traffic, and normal user activities. Exploits are identified by their *signatures* (known patterns of attack). These are the identifying methods that the malware or attacker has used to gain entry into the network. Network analysis software alerts the security team when a signature attack is recognised. That's all well and good for real-time monitoring but it most always means that the deed was done. Cybersecurity has moved on from a complete reaction to activity to one where networks are managed based on risk. Each entity involved in the network's activity is scored based on the risk. An individual can think of this like having a credit score, which is also a form of *predictive analysis.*

Predictive analytics gives an individual a look into the future, although uncertain. One approach, which someone might call an "on the doorstep" scenario is being able to identify an intrusion without having a prior signature. Machine learning in AI actually learns how to recognise patterns far better than a human (Coombs,

2018). By analysing all kinds of previous attacks machines have begun to have a “gut feeling” or predictive ability about what might be an attack, even if it doesn’t match a previously known signature. With the network in constant flux, it becomes a superhuman job to determine exactly what a network’s normal behaviour looks like. There are also malware programs that sit on the network appearing innocuous because the damage (normally data theft) is long term (Coombs, 2018). These are called Advanced Persistent Threats (APTs). They’re cleverly designed to be overlooked by network security programs and to remain in place for as long as possible.

Artificially intelligent machines today are not the sum of their programming as they once were. They analyse great sums of data, the more the better, and find patterns that might have been otherwise unrecognizable. Machine learning may examine millions of math problems and their results and determine, based on a pattern, what the result might be. Applied to cybersecurity the goal is to examine this network data and apply everything it has previously learned to augment a human-led security team.

The acronym URL, not to be confused with the Uniform Resource Locator of the World Wide Web, is an acronym that stands for the following:

- **Understand:** Examine the mass of prior research using NLP. This information can be found within videos, books, magazines, journal articles, and yes, even PowerPoint.
- **Reason:** Provide insights based on analysis that include what type of attack may occur, or may have occurred, and the types of threat entities involved in the attack and their relationships.

- **Learn:** Up to the millisecond research findings continually add to the corpus of knowledge. New insights are continually created based on new information.

2.5.3 Algorithms

Sentimental Analysis is a broader field in text mining which has a great role in text classification. (YooJelnsong and OkRanJeong ,2018; Subramniam et al., 2017; Pandey, Saraswat, 2017). Twitter Analysis is one of the subareas in Sentimental analysis where the tweets are being classified into different categories, Twitter serves as a source for the society in gathering the people's thoughts, often Twitter contributes very high in marketing (Andrea, Ducange, and Renda, 2019). People share their thoughts regarding various products in the market, which may include the quality of the product, the most current trending product in the market, which marks out a varying graph of the thoughts of different people (RatabGull, UmarShoaib, SabaRasheed, and Abid, 2016). Not only in marketing, Twitter also has a major effect in many fields, in many cases the tweets of people change the situation. Analysis over the tweets is necessary since it gives an overall opinion in many cases and gives a clear-cut idea of what various people think in a situation. For the analysis purpose a proper algorithm is required in order to provide accurate results (Prabha, Lakshmi, and Subbulaskmi, 2019).

Within this research there will be three algorithms being used, the research will conduct predictions about abnormal patterns in social media data available on the Internet using machine learning data mining algorithms such as:

- Linear Regression
- FWKNN: Feed-Forward K-Nearest Neighbours Algorithm
- MNB: Multinomial Naive Bayes Algorithm

- FWMNB: Feed-Forward Multinomial Naive Bayes Algorithm
- KNN: K-Nearest Neighbour Algorithm
- K-means
- PCA: Principal Component Analysis

2. 5.3a Liner Regression

Linear Regression, in machine learning there are a set of input variables (x) that are used to determine output variable (y). A relationship exists between the input and output variables, the goal of machine learning is to quantify that relationship. The relationship between the input variables (x) and output variable (y) is expressed as an equation of the form: y = a + bx. Thus, the goal of the linear regression is to find out the values of coefficients a and b. for instance, a is the intercept and b is the slope of the line.

2.5.3b Feed-Forward K-Nearest Neighbours (FWKNN)

FWKNN also known as: Feed-Forward K-Nearest Neighbours Algorithm. KNN is a nonparametric learning method and sensitive to distance function due to the inherent sensitivity of irrelevant attributes. FWKNN modelling is applied which is based on weighting. FWKNN determines the weight of the attribute by identifying the nearest k-neighbour which reduces inherent irrelevant attributes in measuring the distance (Pratama, Tulus and Effendi, 2019). By providing weight to its attributes, FWKNN makes a distinction to the features, meaning the more significant attributes have a higher impact on distance determination this can reduce error in the classification method (Pratama, et al., 2019).

2.5.3c Naïve Bayes

Naïve Bayes, to calculate the probability of a hypothesis (h) being true, given our prior knowledge (d), we use Bayes's Theorem as follows:

P(h|d) = (P(d|h) P(h)) / P(d)

where:

• P(h|d) = Posterior probability. The probability of hypothesis h being true, given the

- data d, where P(h|d) = P(d1| h) P(d2| h)....P(dn| h) P(d) • P(d|h) = Likelihood. The probability of data d given that the hypothesis h was true. • P(h) = Class prior probability. The probability of hypothesis h being true
- (irrespective of the data) • P(d) = Predictor prior probability. Probability of the data (irrespective of the
- hypothesis)

This algorithm is called 'naive' because it assumes that all the variables are independent of each other, which is a naive assumption to make in real-world examples.

2.5.3d Multinomial Naïve Bayes (MNB)

MNB also known as: Multinomial Naïve Bayes. This data mining algorithm has been widely used in text classification due to its computational advantage and simplicity. MNB is a modified form of Naïve Bayes Classifier, MNB is also a probabilistic approach which is like Naïve Bayes (NB). MNB is specially designed for text documents to calculate the occurrence of each word (Prabha, Lakshmi and Subbulaskhmi B, 2019). Naïve Bayes (NB), works based on the conditional probability (considering the conditional independence of the features), while in comparison with Multinomial Naïve Bayes which is based on the multinomial distribution. The

multinomial Naïve Bayes classifier considers the multiple occurrences of each term (Prabha, Lakshmi and Subbulaskhmi B, 2019).

Again, Prabha, Lakshmi and Subbulaskhmi B (2019) suggested, why MNB is used. Naïve Bayes classifier is one of the most widely used method for text mining, Multinomial Naïve Bayes could be said as a upgraded version of the existing Naïve Bayes classifier, and it effectively manipulates the word count by calculating the frequency of each word, whereas in Naïve Bayes classifier the frequency of the words does not have much effect on the working of the algorithm. It is known that the frequency of each text has a higher impact in categorising the text into different categories. Hence Multinomial Naïve Bayes is considered to be best for the purpose of classification of the text.

Prabha, Lakshmi, and Subbulaskhmi, (2019), used this model of MNB and explain the process flow of MNB: and the classification process includes a number of sequential steps, which are as follows:

Step 1. Dataset Gathering:

- the data set required for the project or research that is being gathered.
    - Pre-Processing: the collected dataset is pre-processed, and the unwanted symbols and values are removed so that the dataset is completely fir for the classification process.

Step 2. Implementation:

- The implementation is done by loading different packages required for the algorithm. The dataset is loaded as a csv file and then the steps such as feature extraction, finding the term(s) frequency for each corresponding term(s) is being done and then the algorithm is applied on the processed testing dataset leading to the expected result.

Step 3. Gathering of Results:

- The results are being obtained as two categories as positive and negative which denotes if the corresponding tweet is positive or negative.

Step 4. Performance Comparison:

- The existing work of Naïve Bayes is being compared with the other current proposed methods, thus helping the confusions matrix.

2.5.3e Feed-Forward Multinomial Naïve Bayes (FWMNB)

FWMNB also known as Feed-Forward Multinomial Naïve Bayes. FWMNB is a CFS (correlation-based feature selection)-based feature weighting approach to these naïve Bayes text classifiers. To overcome the shortcoming confronting the multi-variant Bernoulli model, the multinomial model is proposed by capturing the information of the number of times a word occurs in a document. This multinomial model is widely called multinomial naive Bayes (MNB). Jiang, Wang, Li, and Zhang, 2016 suggests: MNB provides on average a 27% reduction in error rate over the multi-variant Bernoulli model at any vocabulary size. However, one systemic problem confronting MNB is that when one class has more training documents than the others, MNB selects poor weights for the decision boundary. This is probably due to an under-studied bias effect that shrinks weights for classes with few training documents.

Finally, Jiang et al., 2016 mentions: although within recent work, supervised learning has shown that these naive Bayes text classifiers, such as MNB, CNB and OVA, have attained remarkable classification performance, all of them assume that all features are independent given the class. However, it is recognisable that the conditional independence assumption in them is rarely true, which would harm their

performance in the real-world text classification applications with complex dependencies among features.

2.5.3f K-Nearest Neighbour (KNN)

KNN also known as: K-Nearest Neighbour Algorithm. KNN classifier is one of the most common and easy to implement classifier in the machine learning domain, achieving competitive results compared with most complex methods, and sometimes it is the only available choice, for example when used for content-based image retrieval (Hassanat, 2018).

However, Hassanat (2018) continues, with mentioning how KNN is a very slow classifier and a lazy learner. For instance, testing any example, the KNN classifier cannot produce a small fixed-size training set of *n* examples, in *d* dimensional feature-space, the running cost to classify one example is O(*n.d*) time, Hassanat (2018) submitted: since we have the blessing or curse of bid data, where n and/or d are relatively large values, big data sets includes their ability to provide a rich source of information to the classifiers for a better learning, while the curse of big data sets includes their very large sizes.

2.5.3g K-means

K-means is an iterative algorithm that groups similar data into clusters. It calculates the centroids of k clusters and assigns a data point to that cluster having least distance between its centroid and the data point. Here's how it works: start by choosing a value of k. for example, use k = 3. Then, randomly assign each data point to any of the 3 clusters. Compute cluster centroid for each of the clusters.

2.5.3h Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is used to make data easy to explore and visualize by reducing the number of variables. This is done by capturing the maximum

variance in the data into a new coordinate system with axes called principal components. Each component is a linear combination of the original variables and is orthogonal to one another. Orthogonality between components indicates that the correlation between these components is zero.

The first principal component captures the direction of the maximum variability in the data. The second principal component captures the remaining variance in the data but has variables uncorrelated with the first component. Similarly, all successive principal components (PC3, PC4 and so on) capture the remaining variance while being uncorrelated with the previous component.

Most of the proposed work in this domain depends mainly on a 'divide and conquer' algorithmic principle, this is a reasonable and rational approach to be used with big datasets. Therefore, most of these approaches are based on clustering, splitting, or partitioning the data to reduce its enormous size to a manageable size that can be targeted later by the KNN. However, such approaches inherit a key problem that is the determination of the best number of clusters, as more clusters means less data, and therefore faster testing, but less data also means less accuracy, as the remaining examples might not be related to the tested example. On the other hand, a small number of clusters specifies many or a vast majority examples for each cluster, which increases the accuracy but slows down the classification process.

2.6. Conclusion and Gaps for Further Study

All the previous studies of social media analytics (data mining-based) that are reported in the open literature focus on cyber fraud, cyber bullies and cyber hate crime. Cyberstalking analytics has not been given great attention by the researchers in the past and this motivated the present study. In addition, lightweight data mining

algorithms have not been used to detect cyberstalking on social media platforms with the use of Twitter (Karyofyllis, 2018). There are many gaps within academia surrounding cyberstalking research, not a lot of research or knowledge of this field; especially with the use of the use of the social media platform that is being used for this study, is not shown within academia. However, there are many similarities to other aspects regarding research that compares to this study.

This research is intended to achieve an extensive understanding of the research questions being asked. This research aims to focus on, identifying cyberstalking, using a data mining method to identify metrics of cyberstalking. Therefore, it will qualify a metrics measurement to determine the threshold value, either with communication or rapport and the linking of different types of patterns from normal online users compared to a potential online cyberstalking user(s). Lastly, the research will apply the proposed model to a real data set on social media. Furthermore, the data analytics will determine whether there is a correlation between cyberstalking and the use of digital technology within social media networks.

The research conducted will help further current academic research on cyberstalking. All in all, this research will pave the way on how cyberstalking is seen within academia and how it is being study or researched. In conclusion, data gathered within this study will inform and pave the way for a new area of communication with the use of recurring themes, negative and positive terms, from within social media network. Finally, the methodology and research proposed will advance current academic research in this field and continue to form an in-depth narration or conversation concerning the profound topic that is cyberstalking.

Chapter 3: Automatic Identification of Cyberstalking on Twitter using NVivo Coding

This chapter involves the preliminary data that was collected for this research. Within a programme that was used to gather the data, that is distributed within this chapter. The programme that was used is NVivo with an extension link called NCapture for Google Chrome. Each programme contributions and assembles the accurate materials from social media outlets such as Twitter, which are needed in this research. Finally, this chapter breaks down into how NVivo works and how it was used for this research. Importantly, shown in the appendix section there are detailed instructions on how to install NVivo on Windows and Mac, as well as all its functions and extensions.

3.1. Introduction

Social media makes use of websites and computer programs to enable people to communicate and share information on the Internet using a computer or mobile device. The information or data published on social media can be searched in search engines such as Twitter using data mining techniques, which helps in better understanding of data for research and organisational functions (Aggarwal, 2011).

Regarding social media, one of the data mining techniques is called NVivo, which can be used to gather data on the topic and current research questions. The NVivo programme consists of codes that mine data from secondary sources, such as literature-based texts, academic articles, journal articles, government websites, documents, and literature on cyberstalking.

Within academia the use of NVivo is not corresponded well, most researchers engaged in qualitative data analysis and have heard of Qualitative Data Analysis Software (QDAS) or Computer Assisted Qualitative Analysis (CAQDAS). It is well known within academia, that NVivo is one of the options for sorting, managing, and analysing qualitative data. However, few qualitative researchers are aware of how long the software has been around or know the ways in which it has been discussed alongside, or within and/ or comparison with 'manual' methods (Jackson and Bazeley, 2019). NVivo, allows researchers to organise and analyse a wide variety of data, including but not limited to documents, images, audio, video, questionnaires, and web/ social media content (Edhlund and McDougall, 2019). In 2019, Edhlund and McDougall described what features NVivo has; the features to use within its programme are vast, here is a list of what NVivo is used for (Edhlund and McDougall, 2019):

- Two software editions: Pro and Plus
- Updated user interface and a new, more comprehensive Navigation View
- Expanded use of smart context dependent Ribbon menus
- Simplified terminology for various functions and tools
- Highly automated multi-language transcript services allowing transcribing of audio and video files from NVivo
- New Crosstab Query for detailed demographic analysis
- Export to SPSS files, SAV, for Classification sheets, Node matrices, and result of Crosstab Queries

When the user has everything set up and installed, he or she can use NVivo and the Google Chrome extension to extract data he or she needs from the Internet. Twitter is a very popular social media platform; it broadcasts its forum or threads all over the world and gains access to many interesting tops. The social media platform that is Twitter is assessed by each programme. Here is a look at how profound Twitter actually is as a social media platform. In 2018, the monthly active Twitter user was 330 million that were posting 500 million tweets per day (Abbass, Ali, Ali, Akbar, and Saleem, 2020). These massive tweets having diverse dimensions of data, used by researchers for different types of inquiries to predict future trends, Abbass 2020 et al., gave pronounced samples as the future marketing outcomes, forecasting box-office movies revenues, flu spreading diseases, disaster response, crime prediction, forecasting election results and so on.

The reasoning as to why Twitter is a great platform to use within social media, because it illiterates how now a day's new wave of social media technologies such as: Facebook, blogs, wikis, microblogging, and Twitter (itself, plays a vital role in formal and informal communications. The microblogging site (Twitter) is an electronic platform where users share their ideas, thoughts, and news in under 280 characters of text (Abbass 2020 et al.,). Twitter is a unique way of following friends and sending tweets (Twitter messages) unlike any other social media networks because the Twitter friendship is not mutual. For example, a person can follow the celebrities without requiring them to follow the person back, therefore, Twitter plays a virtual online world for its users. Virtual world interacts like a real world where the location acts as an intermediate connection (Abbass et al., 2020). The fact that crimes occur everywhere in the world, for upsurge rate of crimes law enforcement agencies are demanding advanced information systems that can help to reduce the crimes and protect the

society. Criminology is the scientific study of crime to find out the causes of crimes by collecting and investigating data. That way Natural Language Processing is a good approach for text analysis (Abbass et al., 2020). NVivo is a vastly new and forthcoming way to conduct data analysis. This programme is extremely useful, records vast amount of data, and helps with analysing data and cost-effective method of research.

The main or prime focus for this chapter is with the use of data mining and analytics on social media to gather the detection of cyberstalking. The problem that is to be solved in the chapter is to identify the possible associate certain unigrams used on twitter and the correlation to cyberstalking. More specifically in detail the question or reasoning as to why this chapter is significant is primarily because this research has never been done before. This however is not considered a problem; therefore, it gave the researcher more opportunity to congregate materials and data to endure the research itself. However, the only problem or issue that was considered is that this research is new and has not been done before, which is how this research would be beneficial to academia. Consequently, the above setback or concern was confronted, by proving that this research would assist academia and overlay the way for cybercrime with the use of social media.

Furthermore, this chapter explores the ins and outs of how the preliminary data was captured, how the data was used and lastly, and how the data was formed. In each section within this chapter, it is segmented by how to gather material, such as the unigrams to use, fellow themes and characters that are frequently used, the social media aspect in terms of cyberstalking, as well as each programme that was used and very beneficial for this study. Additionally, the NVivo and NCapture data that is already in the public sphere is an unlimited tool to use. Whilst the focus is on postings on Twitter this approach includes newspapers, novels, radio, the Internet, or archived

data; these provide valuable learning experiences as data is gathered from each source and analysed within the programme (Jackson and Bazeley, 2019). Pairing these two innovative and unique programmes together benefits the improving narration of cyberstalking within academia. Moreover, each section is accompanied by graphs and charts to illustrate what was prepared and how it was completed. Along with a step-by-step process of how each programme was used is a description of how it works for the purpose of this study.

3.2. Design of The Study

3.2.1 Data Mining Analytics

The intention to the use of data mining techniques on social media is that the data is the enabling factor for advanced search in search engines such as Twitter and helps in better understanding of data for research and organisational functions (Aggarwal, 2011).

Data mining techniques are more than capable of handling the three dominant disputes with social media, which are: size, noise, and dynamism. Each dispute is measured by how the dispute is classified within the quantity of its mention or public perception. Therefore, social media data sets are very voluminous and require automated information processing for analysing it within a reasonable timeframe. Likewise, Adedoyin-Olowe, Gaber, and Stahl (2014) suggests: 'SM (social media) sites appear to be perfect sites to work on especially where opinion/sentiment expression is involved'. As stated above, social media data sets are characterised by the three dominant disputes which again are: size, noise, and dynamism. Analysing these disputes and with the use of data mining on social media can classify the size or volume, noise of postings, and the dynamic of the 'tweets' Aldedoyin-Olowe et al.

(2014) explained: ‘SM is characterised by noisy data such as spam blogs and irrelevant tweets in the case of Twitter. The dynamism in SM data sets are/is versatile in handling such dynamic data’.

The data mining method that is being used here is to measure parameters such as:

- Terms / key words
- Number of postings / conversation or connections
- Probabilities (key words appear)
- Weightings of terms or key words
- Location of postings or connections (IP address)

The method that is being used, is to expand and to test the aim of this PhD research with the use of data mining and machine learning, to have security metrics to detect cyberstalking from social media platforms with the use of Twitter. In addition, with the propose to use towards future cyberstalking suggestive content to make predictions on detecting cyberstalking, that was previously mentioned in the beginning of this chapter.

3.2.2 Experimental Setup

As previously stated, NVivo with NCapture, which is a free extension used with NVivo provided by Google Chrome. Respectively the programmes were used to gather data from Twitter and coded with NVivo. Three threads, Cyberstalking (no # used), Stalking and Fear (no # used), and lastly, Cyberstalking (# used), were originally searched on Twitter; each thread mentioned is explained below in further detail. A frequent word quire was then performed on each Twitter thread and the reoccurring unigrams and keywords were looked over. Once that each thread was

looked through and completed 15 of the commonly used or repetitive unigrams that have been reoccurring or perceived during the procurement the data were then taken. A Twitter search of each unigram was conducted 5000 tweets from each (unigram) search, alongside the assistance of NCapture were then collected. However, before each Twitter thread was captured the authorization to capture the tweets onto the computer; was experienced and then granted by the user using NCapture. Immediately, after that was accomplished each tweet in each thread, that can be linked to cyberstalking was later selected and coded. For example, as you will be shown further on in this chapter, the unigram Follow(s) or Follower(s) out of 5000 tweets, 41 of those tweets can be linked or correlated to cyberstalking.

Whilst using Twitter each unigram search brings up numerous tweets, the use of NCapture, is to collect each individual tweet as a data set from within Twitter. This method was very beneficial for this study making it less time consuming. Rather than obtaining each tweet by hand and coding them by hand as well. Once the data set was loaded and saved, the data set was then upload into NVivo using their NCapture upload tool under the Data tab, in which it automatically asked which data set the user would like to upload. Immediately, the tweets were downloaded into NVivo where they were polished and cleaned up then certain data was removed for sensitive information such as: names, usernames, bios for their profiles, numbers of followers (since that is irrelevant for this study) and so on. Furthermore, while leaving each tweet that was present along with the hashtags, location, web, and the tweets respected coordinates. This process was done, because of the prominence that was needed to make sure anonymity was kept of each person and their tweet. Hence, the reasoning as to why the username, bio, list of followers, retweets, likes, and the category under name was removed. Once the thread was cleaned up and all sensitive information was not

included, each tweet was gone through and tweets with that unigram and any link or correlation towards cyberstalking were coded.

Below are screen shots and/or pictures of how each of the above-mentioned methods was performed. Here is a look at how each programme functions; the unigram “annoying” is the term being used in this screen shot and shown for the purpose of this section. As can be seen within this chapter, multiple usernames and profile photos that are shown on twitter through the example screenshot(s). However, that sensitive information was edited over for the purpose of this segment to show how the program works.

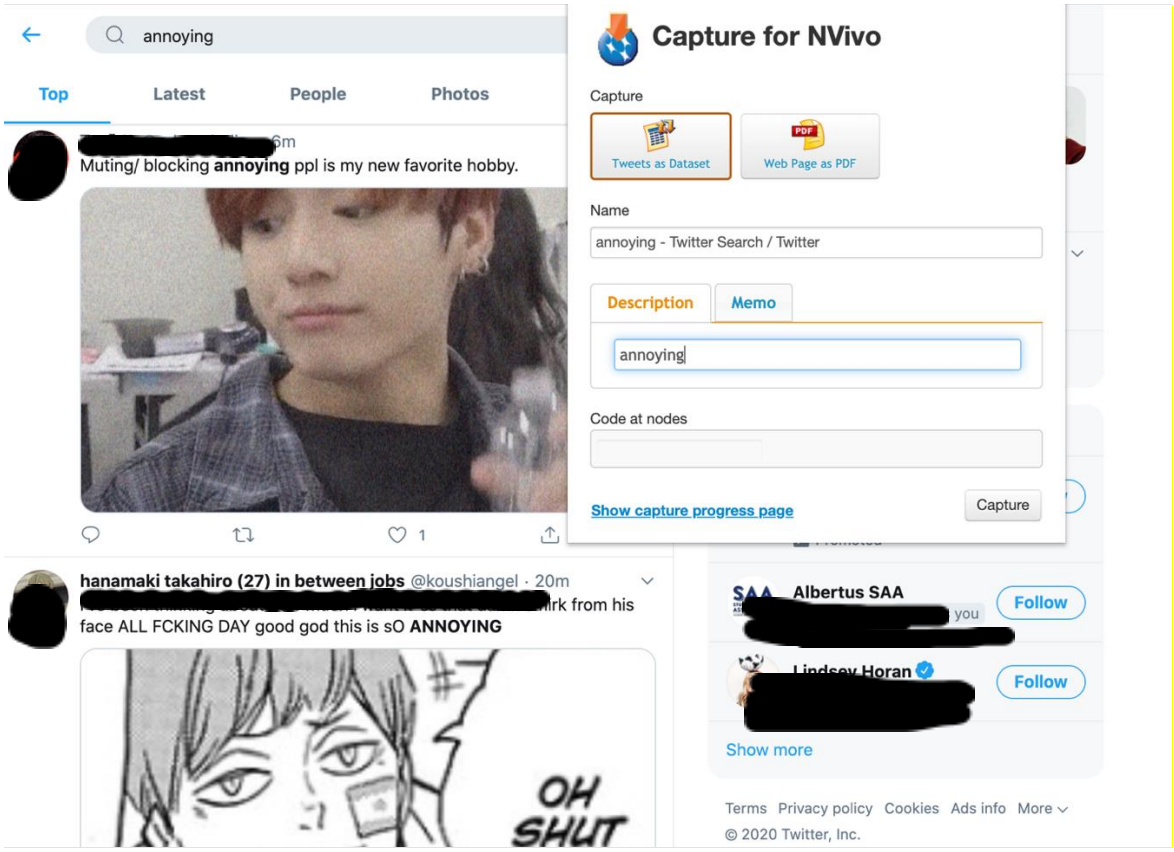


Figure 3. Twitter with the use of NCapture: “annoying”

As shown in Figure 3, all the data and programmes that are being shown and were used were all on Google Chrome. Google Chrome as stated throughout this chapter was used during the entire data collection process. As each of these programmes would work better with Chrome. Although, if a Mac were to be used,

Google Chrome would need to download on to that device. For instance, a Mac device was used for this entire study. For a Windows computer or laptop, Internet Explorer would work best. Nevertheless, that does not mean a Google computer is a must, any type of PC or laptop could be used, as the researcher stated and has shown. Although there needs to be measures to make sure each software works or is compatible on the PC or laptop chosen. As well as, if the free trial of NVivo were not efficient for the proposed research at hand, then the licensing should be bought, and the activation process must be enabled since it is a requirement to use the software programme. Furthermore, Figure 3 shows a Twitter search of the unigram “annoying” below the search are numerous tweets with the term “annoying” being used. The entire feed was collected as a data set or PDF file to be uploaded to NVivo, either could be upload according to the user’s personal preference or what type of data set to be used. PDF files capture the entire webpage or article the user was looking at as the webpage or article itself. Whereas the data set captures the information or data as a set to use within NVivo. For the main purpose of this study PDF files were not used in the data collection process.

The data for this search thread was collected using NCapture provided by NVivo. NCapture as previously stated is a Google Chrome extension found on the right-hand side of the tool bar. Once it was selected the little window to the right of the screenshot popped up on the feed. “annoying” was typed in the box labelled description for this purpose; however, it could be labelled anything that is significant to the research. There was an option to save the entire thread into a node or code it, for this study that function was used, and each thread was gone through individually and the findings were then coded as a node separately. After labelling and deciding

how to save the data as a data set or PDF, the capture button on the bottom right was clicked to start capturing.

Now, after the ‘capture’ button was clicked another widow appeared on the screen stating what is shown in Figure 4. This screenshot is the user giving NVivo / NCapture the authorization to access the twitter account to obtain the information the user would like to gather and use. Once the user clicked ‘Authorize app’ the feed began to be captured. If NCapture were not authorized to use the account, then it would not be able to gather the data from Twitter. So, once accepted and proceeded the next page showed as bellow.

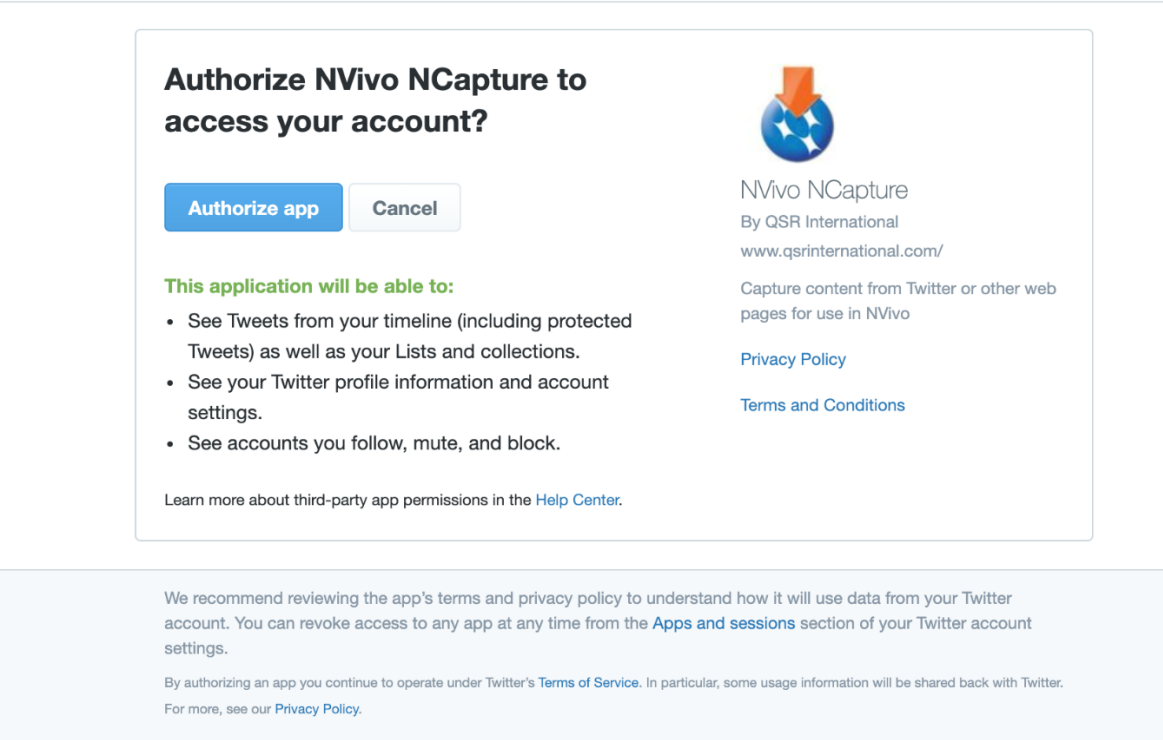


Figure 4. Authorization NVivo NCapture to access a twitter account

Beneath in Figure 5, which shows the number of tweets captured. As can be seen it shows the thread which was named annoying, and the type of data source which was tweets from Twitter. Also, the messages in this case were 1500 Tweets captured and the status was still loading; however, this process could be stopped whenever seen fit for the research or the induvial purpose. Please note that if the

capture was not stopped manually for the purpose of this research. The more threads that would have been captured would have been greater than needed and harder to manage how many threads that were needed per unigram, key word, or data set. It is very imperative to know that NCapture only captures a certain amount or number of threads as a data set at a time; if that is the case a notification bar at the bottom of the screen will pop up and inform the user to wait a few minutes and try again later. Therefore, the machine as well as NCapture and Twitter need a break to not let the user capture more threads than Twitter allows.

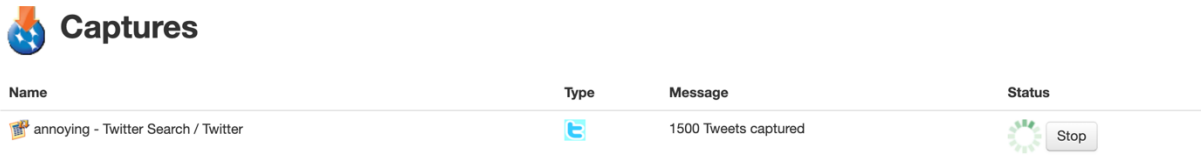


Figure 5. Number of tweets captured while using NCapture

Moreover, shown below in Figure 6, the NVivo project twitter.data (which was previously saved on the computer to be used) opened and ready for the data set that NCapture had captured from Twitter. In order to open a data set in NVivo, a new project was created and opened or one that was previously saved beforehand, e.g. twitter.data, was already opened and previously saved beforehand. As can be seen on the top of NVivo it has a toolbar and under that it has:

- Project - for a new project
- Documents - for all documents that need to be uploaded into NVivo
- PDF files - such as articles, journals, social media, and news reports
- Dataset - for data set files
- Picture - for importing any picture into the project

- NCapture - which is used for what is being shown in this section
- Codebook - which allows the user to export a codebook to a desktop if codes have been saved under nodes

The tool bar has many functions that can be used while examining the data within NVivo. As can be seen, above the list that is mentioned above there are more options to use or see that NVivo offers. Which are:

- Home - is all of the function within the software
- Data - click to import the data as can be seen below the ways to import data
    - Documents, PDF's, Dataset, Picture, and NCapture
- Analyse - to tools to analyse the data that have been imported
- Query - is the functions that can be run within NVivo
    - Text search, word frequencies, coding, matrix coding, crosstab, coding comparison
- Explore - is where the data is graphed, and other options can be used to elevate the data
    - Mind map, concept map, charts, hierarchy chart, explore diagram, comparison diagram, files classification sheets
- Layout - the layout of NVivo and or the home screen
- View - the way the user views the layout in more detail

However, even though NVivo has license for students, it is imperative to note that some tools are not authorised to be used within that licensing. For example, certain features will not be able to use unless the person or individual pays for a premium service for NVivo. Nevertheless, the student account and service are more than enough for this study or researcher and does not hinder the process. As stated before, it is essential to note that for future projects or research, if other tools or services are required and the student or free version of NVivo does not suffice one of the premium packages may.

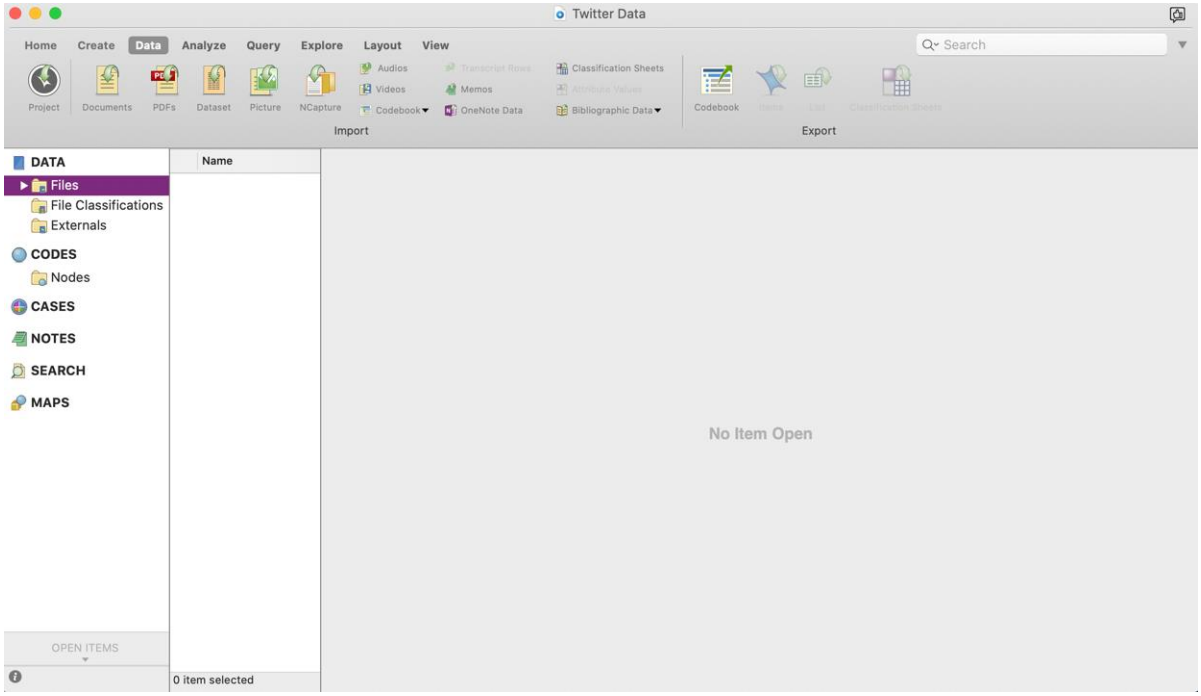


Figure 6. NVivo Programming interface user account example

Displayed in Figure 6, helps bring forward what appeared after clinking on NCapture to upload the data set. A window of all the data sets that were already captured and imported appeared, the newest one was highlighted; not having a check or tick mark next to it means it needs to be imported. In addition, more than one file can be imported at the same time, and it could be an older file or a newer one as well. The import button was then clicked to load the data set into NVivo. As Figure 5 shows,

‘Files” was highlighted, now that was where this data set was uploaded too, but that can be changed if desired. Likewise, after clicking files, if there were different folders for each unigram the user could upload the data set straight into the folder, the user would like to have the data set in instead of using the file folder.

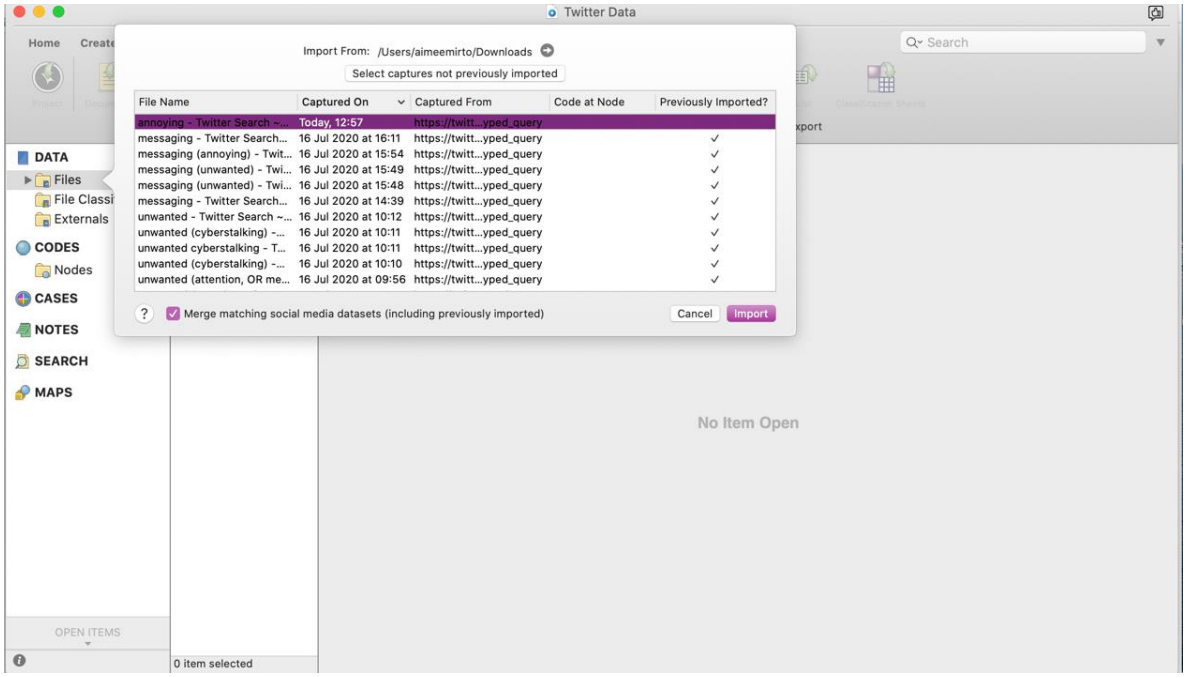


Figure 7. Importing the tweets collected using NCapture into NVivo

Exhibited in Figure 8, which is underneath this section, shows the data set loaded into NVivo edited to the user’s preference and cleaned up. The data set was edited and cleaned up manually before this screen shot was taken. Conversely, if the user wanted to edit and/or clean up the data according to his or her preference. Additionally, in order to do so if the user looks at the bottom of the figure the user can see a section that says columns which is under the number eight in this screenshot. If the user clicks the (columns tab) the user can clean up the data to meet the user’s needs or requirements. As previously mentioned, the username, bio, list of followers, likes, and tweet ID were removed from this data set to keep the anonymity of each

person or individual and their tweet. Once that was done for instance in this case the researcher was left with everything that is seen on the screenshot below. When the user is at this step, then the user could go through tweet by tweet and coded any tweet that was beneficial for their research. For example, if tweet number 5 the one that was highlighted (as seen below) was constructive for this research, the user would then highlight the text, go to the right of the screen, and click node and add that to its appropriate node and then that tweet is coded. In addition, if the user thought that tweet would be beneficial in another node as well as the one the user previously picked, the user would select the entire tweet, right click, and select add to existing node and a list of all the nodes would pop up and the user ticked the ones the user would like that tweet to go into.

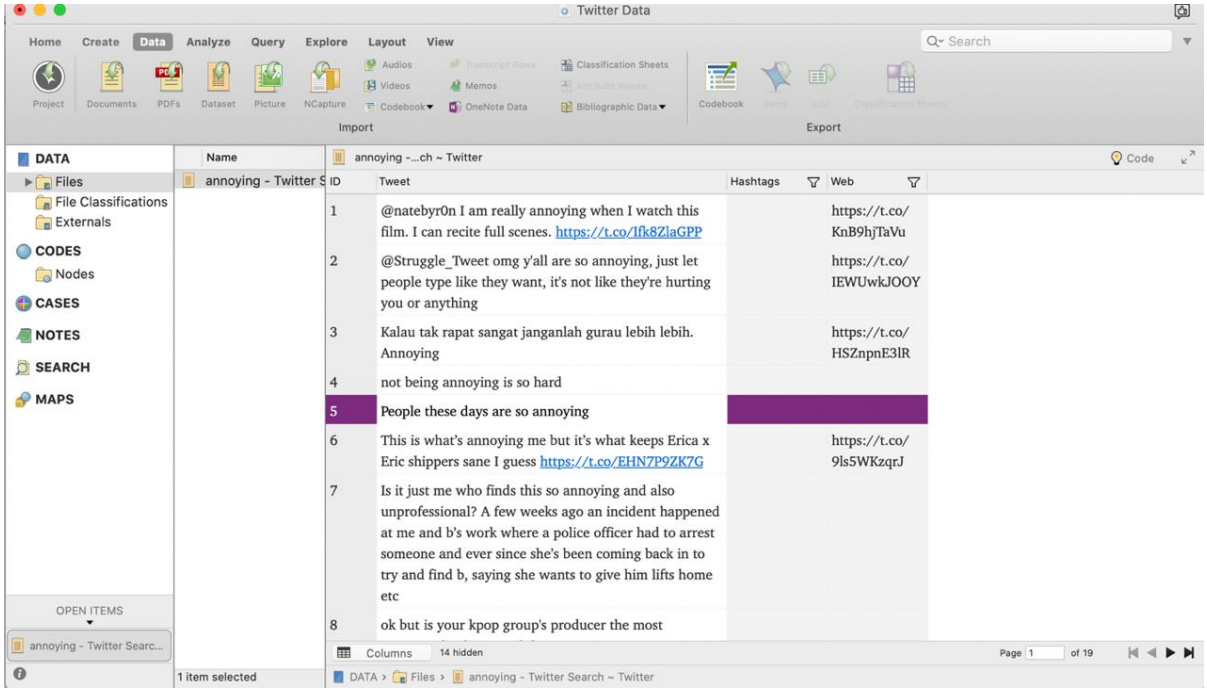


Figure 8. Data Set created from tweets collected shown in NVivo

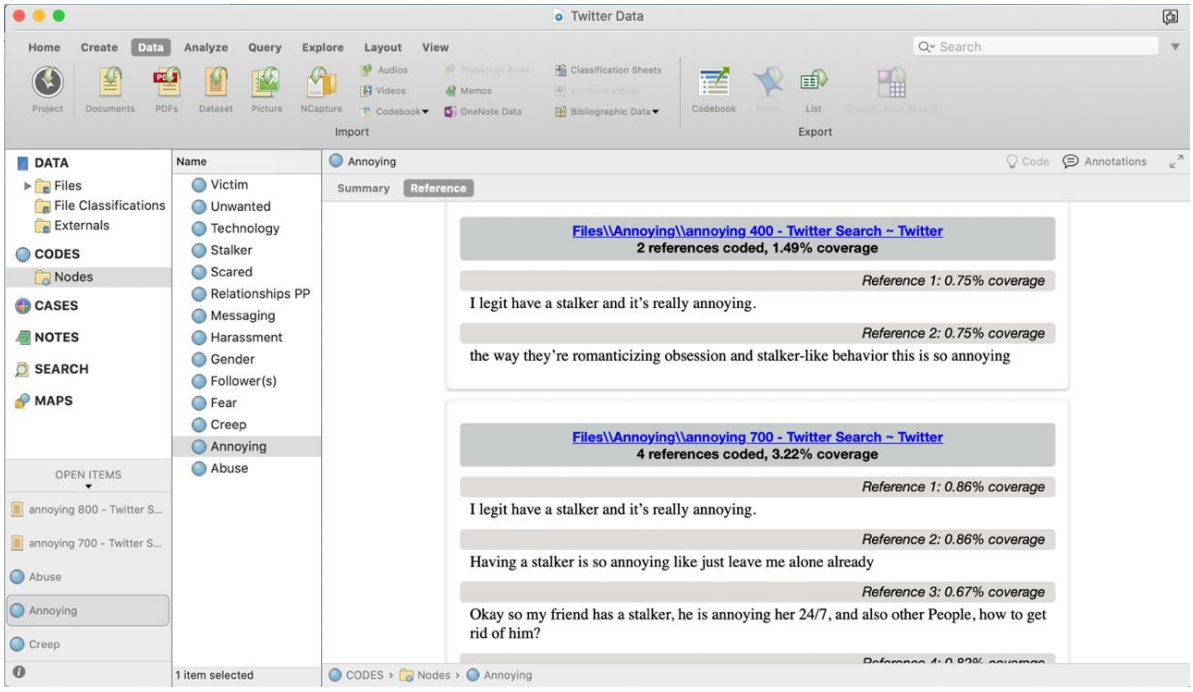


Figure 9. Data Set created from tweets collected shown in NVivo Continued

The above Figure, Figure 9 illustrates the nodes and the references that were correlated with each node or unigram. As previously stated for Figure 7, tweet number 5, which highlighted in the Figure 8 was taken, because it was constructive and beneficial for the research. Once the tweet was considered valuable, the text was highlighted again as stated before, the user would go to the right of the screen and click node and add that to its appropriate node and then that tweet was coded. After the tweet was coded into the node the user wished to use, this screen appeared showing all the user's tweets or reference that were correlated with those unigrams. As can be seen below 14 out of the 15 unigrams were currently coded within NVivo while using Ncapture.

Twitter: Unigrams

Nodes

Name	Description	References
Abuse	Tweets with the term abuse	9
Annoying	Tweets that use “annoying” could be linked to cyberstalking	4
Creep	Tweets that mention creep linked to cyberstalking	51
Fear	Tweets about potential fear of cyberstalking -living in fear of cyberstalking	17
Follower(s)	Tweets about follower or followers or following being compared to cyberstalking	24
Gender	Gender of who the tweet is about	15
Harassment	Tweets that use Harassment	12
Messaging	Tweets with messagings and could be linked with cyberstalking	26
Relationships PP	Tweets about past pr present relationships that can be linked to cyberstalking	21
Scared	Tweets with scared linked to cyberstalking	1
Stalker	Tweets linked to how stalking a can be done on twitter	55
Technology	Use of technology	7
Unwanted	Tweets linked to unwanted and cyberstalking	19
Victim	Tweets associated with victim	2

Aug 5, 2020

Figure 10. Code Book detailing tweets collected and their relationship towards cyberstalking exported from NVivo

This codebook presented overhead in Figure 10 helps explain how the codebook is outlined. When the user used nodes to help organise the data within NVivo, the user could break it down into the name of each node which the user could see in he left column. In the middle column the user could see the description of each node (which the user added and can be described to the user’s preference) and how each compound was dived into their own sections; and why each tweet was selected and put into that node itself. There would normally be a third column that would indicate the total number of point of files (chain of tweets in this purpose) for each node for example how many files there were in each node. However, that section was not needed that for this research since the user knew he or she was using 5000 tweets

each therefore that section or column was deleted entirely. Lastly, on the right are the references this column shows how many tweets were associated to the unigram itself.

These next few figures assist the illumination of each Twitter thread that was put into in NVivo and how they were run through the word frequency tool or function. The instructions of how each thread was brought into NVivo is the same as mentioned above in the previous figures. They help explain how to gather data from Twitter using NCapture and imported into NVivo. As seen in Figure 9, the thread was uploaded from Twitter into NVivo using NCapture. The thread itself did not need to be cleaned up in comparison to the others unless the user personally would want to clean it up. Once the thread was in NVivo and the file was to the individuals liking and if the user wanted to run a word frequency like the researcher did. The user would click on query within the tool bar and then word frequency. After clicking on word frequency, the user was brought to the next figure. In Figure 10, the functions to run the word frequency are as follows:

- Select the filed the user wishes to run the word frequency on
- Once selected, the user can choose if the user wants to have the frequency to be an exact match or have stemmed words. For example: talk
    - Stemmed for example would be talking
- Then the user can adjust the minimum length of the word to the preference
- Then the user has the choice to display all the words or x (up to 1000) most frequent words

When the users are finished adjusting with the settings and functions until the he or she are happy with what the possible outcome the query can be run. After the query

ran all of its tests and finished the user can export any and all the document(s) from NVivo to a desktop and opened it in a word document, to do more editing if the he or she chose to do so. Moreover, the user could also open it into an Excel document and edit, graph, or chart the data as the user please

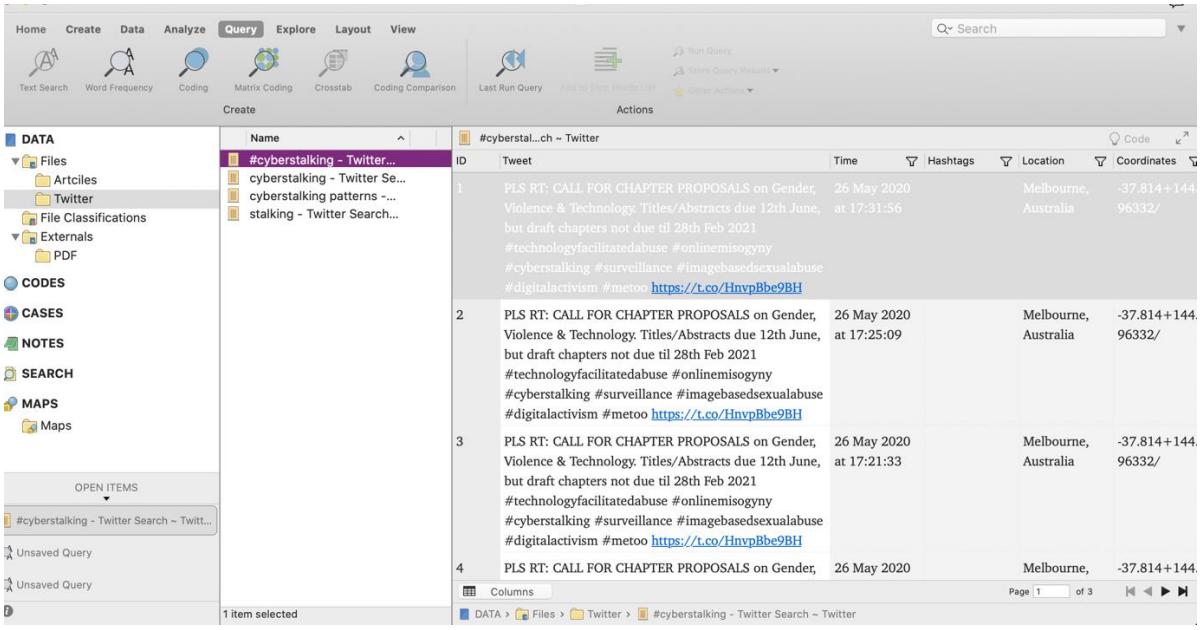


Figure 11. #cyberstalking thread in NVivo

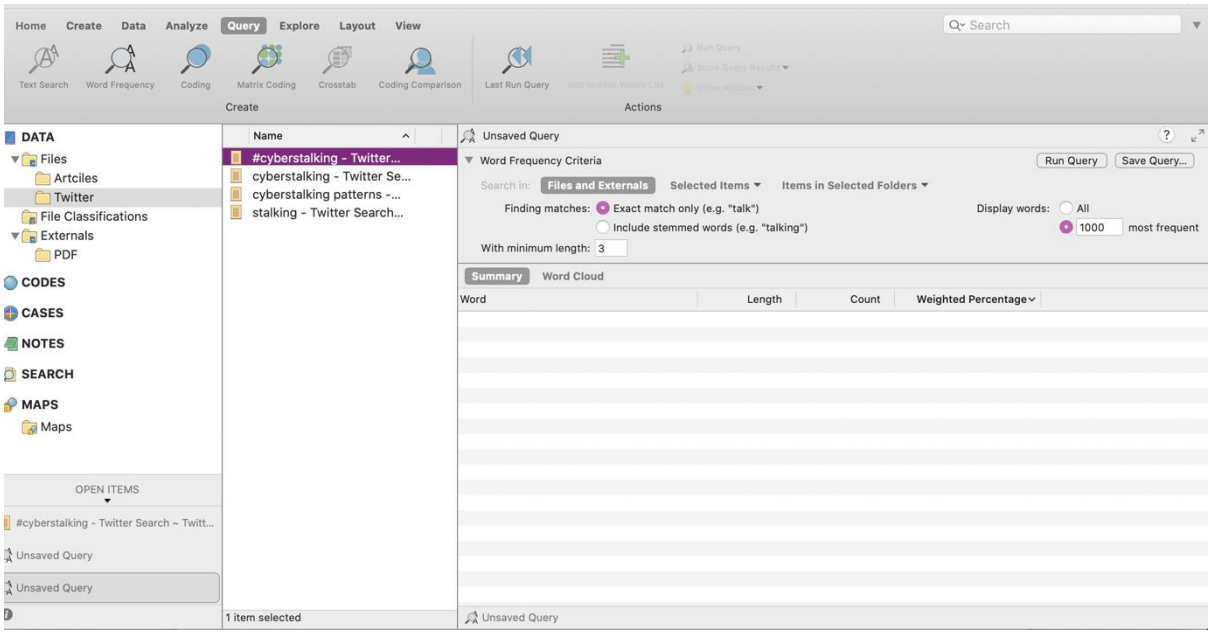


Figure 12. Word Frequency for #cyberstalking

In Figure 13, the user could see the finished word frequency that was just conducted. In the figure below it shows the entire word frequency for the thread #cyberstalking. The user noticed the selected file that was chosen for the programme to run this function, which in fact was #cyberstalking-Twitter. If the user looked at the summary that was placed before him or her, the user would notice five different columns that were listed the columns show: the word, length, count, weighted percentage, and the similar words that were listed or used in the text. Under the word column it expressed the list of all the frequent words used within the thread the user was about to run the word frequency on. The length shows how many letters or characters make up the words that are used within the thread. Length indicates the length or character number for each unigram, for example, the first unigram in the list is: “chapter”. “Chapter”, has 7 as its length meaning the unigram chapter, has seven characters to make up the word. Moving onto the next column, count illustrates the total count that the word has been used or seen within the thread. Weighted percentage displays the percentage of how much the word is used throughout the thread. Lastly, the similar words column indicates all the similar variations of the word itself, for example #cyberstalking would be: #cyberstalker, #cybersatalkings, #cyberstalk and so forth.

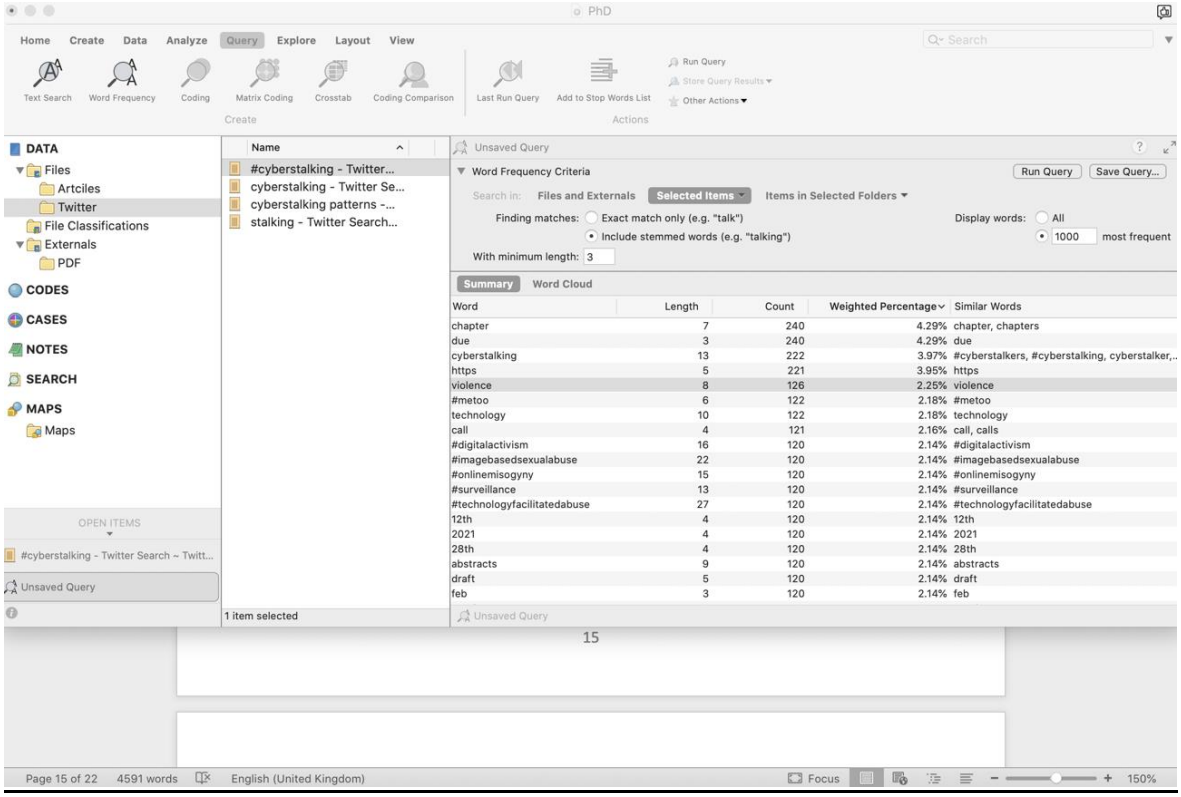


Figure 13. #cyberstalking in NVivo

Some of the precautions that were taken against likely sources of bias were how to keep anonymity of the individuals' social media account, which the researcher was using on this occasion, Twitter. Another precaution that was used was on each tweet itself, if a tweet was thought to be more demographical and not well represented for this study then the tweet itself was discarded. As well as, if it came across anything that sounded or looked criminal, then the correct authorities would be informed to deal with that matter. The limitations that are inflicted within this study would be in within the design of the study and time efficacy would be a major limitation on this study. Mainly, because of how time consuming it is to gather material on each unigram and sift through each tweet. Therefore, that is why only 5000 tweets per unigram, and not more were used. Another unique issue not necessarily a limitation per say, would be that this study is fairly new and has not been studied before in this manner.

Nonetheless, that does not indicate a limitation as previously stated, hence the rains on this study are more strategic since it has not been conducted and is a vastly new emerging topic.

3.3. Results and Discussion

This section discusses in detail within four sections, the preliminary data set that was collected for this study. During the collection of the preliminary data set, as a reminder a data mining programme called NVivo was used. Ncapture, which is an extension from the Internet that is used mainly with Google Chrome, was also adopted. With Ncapture, data sets were taken and downloaded from Twitter and imported it to the desktop. Once retrieved the document was opened in NVivo, and the data set was edited to what was needed from each set for the purpose of this study. In addition, the newly edited data set was exported from NVivo, saved to a desktop and opened in Microsoft Excel. Accordingly, while Excel ran the new project, each data set was correlated by key words, numeric count of each key world, and the weighted percentage, as well as reoccurring similar words.

The preliminary data set that was collected from three different Twitter feeds which were searched for on the social media platform. These three Twitter feeds helped correlate which unigrams were used to conduct the test of the preliminary data set. For example, comparing each list to one another and viewing the reoccurring themes or key words enabled the resecher to decipher which ones were the best to use for this study. The three feeds that were coded with the use of NVivo are as follows under the subheadings that were used and are listed below.

- 3.3.1: Cyberstalking (no # used in this search)

- 3.3.2: Stalking and fear (no # used in this search)
- 3.3.3: #Cyberstalking (# was used in this search)
- 3.3.4: 15 unigrams against 5000 (tweets each)

3.3.1: Cyberstalking (no # used)

In the appendix section shown in table 1. On Twitter cyberstalking was searched on without the use of a hashtag (#). The main purpose the hashtag (#) was not used in the search, in this primarily was to see how the term searched or flagged alone on Twitter with nothing attached to it. After the search was rendered, 100 tweets that referred to cyberstalking were gathered and NCapture was used to take the data set from Twitter and imported into NVivo. Once each data set was loaded and edited to remain anonymous and not implicating the users from Twitter. After completion, a word frequency test was performed on each set, on the 150 most used words or phrases. The reasoning as to why the word frequency was only set to 150 and not more, was because of the time restrictions of the study. Moreover, a word frequency could be run on any data set up to 1000 most frequent words. However, once the word frequency was finished, the smaller or filler words used were revised and edited out, such as: “it”, “is”, “the”, “and”, “I”, “was” are some of the examples that were deleted since they are not relevant for this study. As can be seen in the appendix section the table shows: key words, count (how many times the key word was used), weighted percentage, and lastly similar words. For this feed that is shown from the Excel spreadsheet in Table 1. It is prevalent that all the reoccurring themes or key words that are being used within relation to cyberstalking were found in order to create this study. As well as, how cyberstalking stands out in comparison with other cyber-crimes and activities on the social media platform that is Twitter.

Furthermore, Figure 14 shows the reoccurring themes or key words in a different light compared to Figure 10. The graph below highlights the key terms or the unigram that are frequently being used in this thread. The thread that is being used for this graph is the same as above cyberstalking no (#) used. However, a much closer analysis of the terms or unigrams shows the most focused term or key word being used are; cyberstalking, stalking, harassment, sexual, man (gender related), video and so forth. Meaning these terms are in correlation with tweets alone to a cyberstalking search in comparison to tweets of each unigram itself without having an advanced search or hashtag in correlation to the underlying search of cyberstalking.

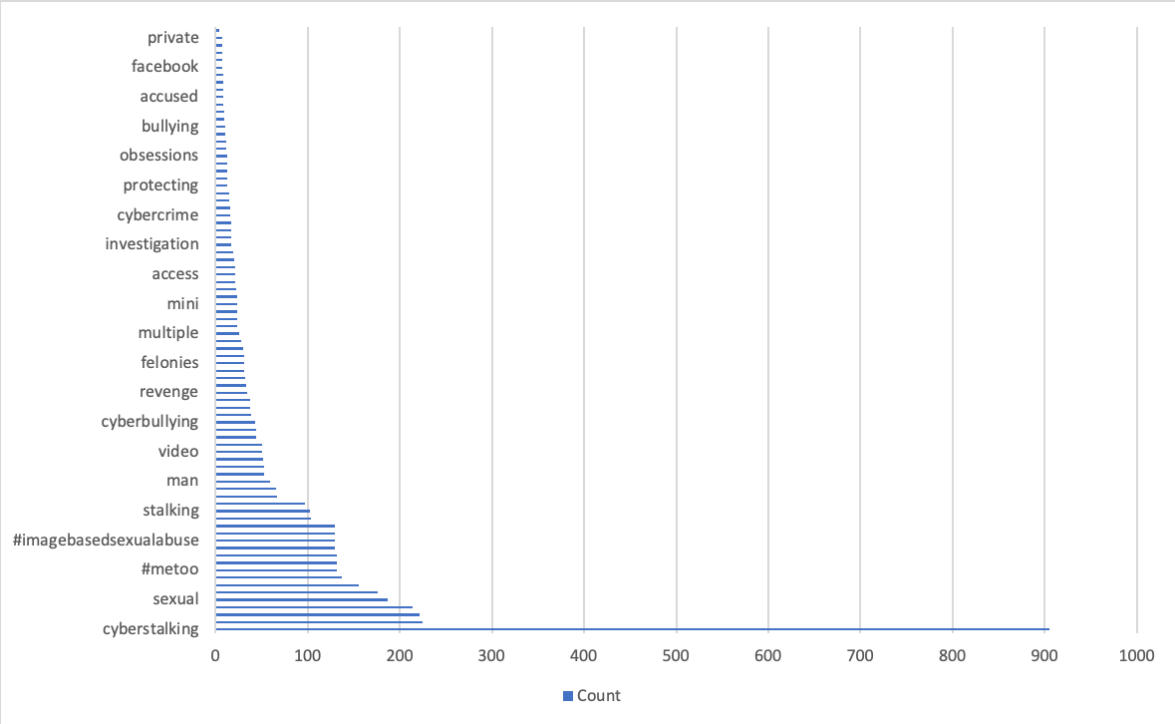


Figure 14. Cyberstalking (no # used) charted from data collected

3.3.2: Stalking and fear (no # used in this search)

Additionally, as shown in the appendix in Table 2, it is the same spread sheet from Figure 13, shown above with the list of: key words, word count, weighted percentage, and lastly again similar words used. Within this data set from Twitter

captured with the use of NCapture containing the thread of stalking and fear, no hashtag (#) was used as well in the search. This search encloses cyberstalking related keywords and terms in correlation to the thread itself. Again, with this data set it ran through a word frequency which took the most common 150 words from the thread imported into NVivo. Likewise, as mentioned earlier the same small phrases or words such as: it", "is", "the", "and", "I", "was" are some of the examples that were deleted since they are not relevant for this study. The reoccurring themes or key words in this thread, as can be seen, were all correlated to stalking and fear. However, there are several vast tweets that were not only linked to stalking and fear; there are various tweets that had no correlation to cyberstalking at all. By conducting the preliminary data, the key finding is that the immense number of tweets were not correlated to topic at hand, which suggests the hashtag (#) tool was probably more beneficial rather than a possible hindrance. However, proving that this outcome was the interesting motivation as to the rational for why again no hashtag (#) was used in the search, which was to see how or if the terms themselves had any parallel to cyberstalking. In comparison to these two threads at hand had many relations to cyberstalking from the massive majority of the reoccurring themes and keywords. Although, there are many differences as well, in association to using a hashtag or not using one during the searching process.

3.3.3: #Cyberstalking (# was used in this search)

In the last thread that was conducted for this study the Twitter function of a hashtag (#) was used in the search. Therefore, the unigram Cyberstalking was search with the function use of a hashtag. The main purpose that the hashtag function was used in any search on Twitter, is everything that is searched under that one term is

only in relation with that term itself. For instance, when #cyberstalking was searched all the tweets in the thread were only related to cyberstalking and nothing else. As seen in Table 3, shown in the appendix, which is the complete word frequency for the search of #cyberstalking. It shows the similarities between all three threads as well as the vast various differences. In addition, the two that are astronomically the same are very different in their own way.

For example, the key terms that are the central focus in all the three threads are: stalking, violence, fear, message(s), annoying and so on. The vast differences are more geared toward the words or unigram searched. For instance, if a broad term like love is searched with no hashtag used, all the tweets that are shown in the thread are related to anything within the unigram love and then some. Therefore three different searches were transmitted and two of them were the same term, cyberstalking and #cyberstalking one with no hashtag being used and the other having the hashtag used in the search. To compare these three threads together, it needed to focus on which unigrams to use to conduct this study.

3.3.4 15 unigrams against 5000 tweets each

Throughout the preliminary data set collection process, many similarities within the Twitter threads, such as keywords, themes, and associations amongst cyberstalking, were noticed. Whilst obtaining the data collected, the most used unigrams that were taken and there could have been more than what was used in this case; however, the total was fifteen of the most used unigrams from each thread and then 5000 tweets were composed for each unigram. The list of the selected unigrams that were used are as follows:

- Abuse

- Annoying
- Creep or Creepy
- Fear
- Follower or Follows
- Gender
- Harassment
- Messaging
- Relationships P/P
- Scared
- Stalker
- Technology
- Unwanted
- Victim
- Violent

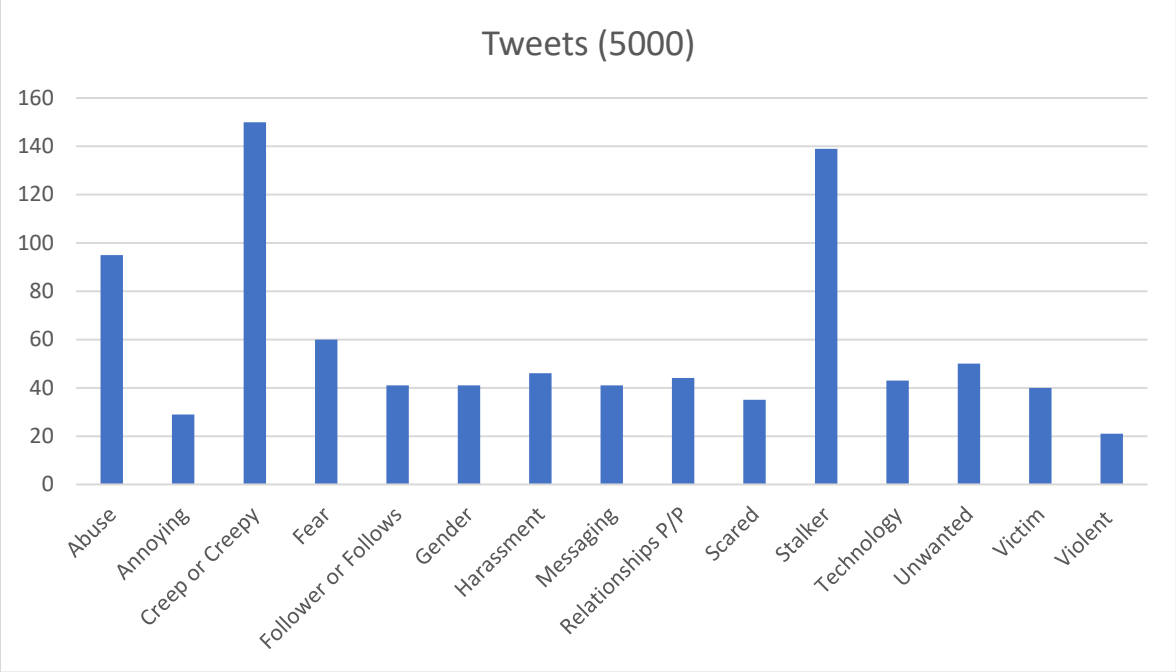


Figure 15. Twitter Data: tweets (5000) collected showcasing the unigrams used for this study

As you can see each unigram shows a significance within corresponding to cyberstalking on the social media platform Twitter. Each unigram that was selected showed a correlation or link to cyberstalking from the preliminary data collected. Each unigram that was selected are not only too general or common but are associated with cyberstalking. A Twitter search was run for each unigram and 5000 tweets were then gathered, i.e. 100 tweet thread at a time and coded each tweet. The tweets that were flagged if they had any correspondence or had any relationship to cyberstalking. They were then coded into a node barring a connection to with that unigram and cyberstalking. Likewise, all of the 5000 tweets were examined, again a thread of a 100 at a time for each unigram manually before coding them into nodes. As previously stated, if the tweet itself were used for more than one node, the tweet would be saved in any of the nodes that is represented.

Unigram	Tweets (5000)
Abuse	95
Annoying	29

Creep or Creepy	150
Fear	60
Follower or Follows	41
Gender	41
Harassment	46
Messaging	41
Relationships P/P	44
Scared	35
Stalker	139
Technology	43
Unwanted	50
Victim	40
Violent	21

Figure 16. Unigram and Tweets (5000) Data Set

As you can see from the above Figure 16, data set each unigram is listed followed by the number of tweets that were captured and considered imperative or associated to cyberstalking. Again, each unigram was not used with a hashtag or an advanced search. The unigram was searched on its own with no connection or manipulation from the term cyberstalking. Respectively, the unigrams that were used for this study were selected from the reoccurring themes or keywords from the above-mentioned Twitter threads in this chapter that were used. These group of unigrams were continuously shown or flagged up on each twitter thread consecutively. The most intriguing finding is that the unigram creep or creepy has 150 tweets in correlation to cyberstalking, whereas violent only has 21 tweets that can be linked to cyberstalking. Each unigram had a vast majority of tweets that were linked to all other outlets or sources. It is very interesting to view how each term weights against cyberstalking rather than standing alone. The findings show that the main four unigrams to focus on would be: creep or creepy, abuse, fear, and lastly the most obvious one stalker. However, that does not mean the other unigrams are not as important; each unigram

brings forth new development on how the term itself is flagged up in correspondence to cyberstalking.

3.4. Summary

In conclusion to summarise, this chapter describes the research methodology used to collect and analyse the data required to address the research questions and to test the hypothesised relationships developed in this study. The chapter begins with a discussion of the research design, followed by the programme in which data was collected and the approach to using the data towards the next selection or chapter. The chapter then continues with descriptions of how the data was collected, the data measurement, and lastly the unigrams preferred for this research. Next, the reasoning as to the choice of methods and data collection and analysis are discussed. In addition, the limitations and recommendations for future studies are suggested and mentioned, as well as what are expected to achieve and potentially continue to do after this study is concluded. Finally, the next step is to explore the data collection and analysis methods, using algorithms and focusing on the metrics of cyberstalking.

Chapter 4: Twitter Data Analysis with the use of R Programming

4.1 Introduction

This chapter presents the data collected from the previous programme that was used and mentioned in the preceding chapter: NVivo and NCapture. As well as, a new programme, R Programming Language will be used in this chapter. The dataset that was used and again mentioned previously, had various amounts of twitter threads that are consumed with endless tweets. Once the data was collected and revised it was then run through R programming and coded further for this project. As mentioned before in Chapter 3, there are fifteen unigrams that were selected to be the main or prime focus in association to cyberstalking on Twitter. Each unigram was the emphasis of 5000 tweets and linking any connection towards cyberstalking. The focus of this chapter will be on R programming Language and how this programme examined and transcended the data collected for the purpose of this research. As well as a random sample data set being introduced in this chapter. Throughout this chapter, each section will go into further detail with regards the Twitter threads and unigrams that were collected and the correspondence or relationship between the datasets being used.

4.1.1 R Programming/ R Studio

4.1.1a R Programming

The perceptive for the use of R Programming for this study or research is beneficial because, R programming is being used within academia to showcase further findings and research for the academic community. R Programming is a remarkable tool that helps correlate the task at hand. For instance, from the R language or programme site, r-project.org (2020): R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S

language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R. Moreover, R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, and more) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an open-source route to participation in that activity.

One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control. R is available as Free Software under the terms of the Free Software Foundations's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. Everything that R includes is as follows:

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent, integrated collection of intermediate tools for data analysis,
- graphical facilities for data analysis and display either on-screen or on hardcopy, and

- a well-developed, simple, and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

The term “environment” is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software. In addition, R, like S, is designed around a true computer language, and it allows users to add additional functionality by defining new functions. Much of the system is itself written in the R dialect of S, which makes it easy for users to follow the algorithmic choices made. For computationally intensive tasks, C, C++ and Fortran code can be linked and called at run time. Advanced users can write C code to manipulate R objects directly.

Lastly, many users think of R as a statistics system. The programmers (we) prefer to think of it as an environment within which statistical techniques are implemented. R can be extended (easily) via *packages*. There are about eight packages supplied with the R distribution and many more are available through the CRAN family of Internet sites covering a very wide range of modern statistics. Likewise, R has its own LaTeX-like documentation format, which is used to supply comprehensive documentation, both on-line in several formats and in hardcopy. Below is the timeline for the extension of RStudio, used with R Programming.

4.1.2 RStudio

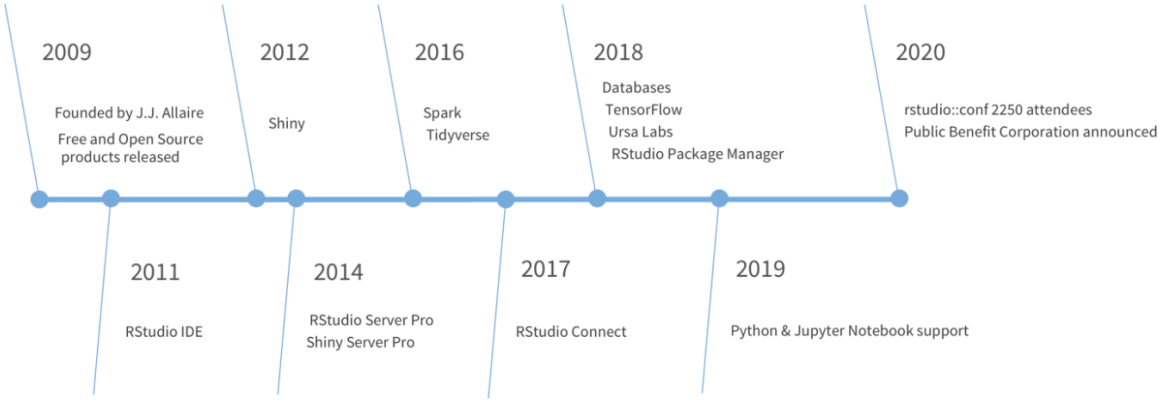


Figure 17. RStudio information and background timeline

The above figure is the publicised timeline for RStudio from their website. RStudio is an integrated development environment (IDE) that allows you to interact with R more readily. RStudio is similar to the standard RGui but is considerably more user friendly. It has more drop-down menus, windows with multiple tabs, and many customization options.

R Programming is inspiring people all over the world. These people are turning to R, Python, and other open-source programming languages, to make sense of all collected data. RStudio, which is inspired by innovators in science, education and academia, government, and industry; RStudio develops free and open tools for R, and enterprise-ready professional products for teams who use both R and Python, to scale and share their work.

In addition, RStudio is in fact being used and downloaded by many. Today, millions of people download and use RStudio open-source products in their daily lives while thousands of organizations and individuals, who have the need and ability to pay for our commercial products on premises or online, to help us to sustain this work. In addition, from RStudio website (2020): it is exciting to consider that we are helping many participate in global economies that increasingly reward data literacy.

This mission statement is from the RStudio website (2020): “The mission statement for RStudio is as follows: RStudio’s mission is to create free and open-source software for data science, scientific research, and technical communication. We do this to enhance the production and consumption of knowledge by everyone, regardless of economic means, and to facilitate collaboration and reproducible research, both of which are critical to the integrity and efficacy of work in science, education, government, and industry. Furthermore, RStudio also produces RStudio Team, a modular platform of commercial software products that give organizations the confidence to adopt R, Python and other open-source data science software at scale - for the benefit of many people, to leverage large amounts of data, to integrate with existing enterprise systems, platforms, and processes, or be compliant with security practices and standards - along with online services to make it easier to learn and use them over the web.

Together, RStudio’s open-source software and commercial software form a virtuous cycle: The adoption of open-source data science software at scale in organizations creates demand for RStudio’s commercial software; and the revenue from commercial software, in turn, enables deeper investment in the open-source software that benefits everyone”.

4.2. Twitter Data Handing in Each Programme

4.2.1 Excel and NVivo

As formerly mentioned in Chapter 3, fifteen unigrams were used as the prime focus in association with cyberstalking on Twitter. The figure shown below which is to remind the reader, about the preliminary data set that was used in R programming to analyse that data even further. Showing the name of each unigram, the reference it

was expended for or the correlation and lastly the number reference itself. However, the importance of each unigram and the data that is balanced along with them is imperative for this research.

Shown in the two figures below, the first figure is the code workbook from NVivo, to remind the reader. This workbook is broken down into the unigrams, the description that the researcher used for each unigram. The references which are regarding the number of tweets that were used as nodes for each unigram. The second figure is the dataset that was made in Excel from the above codebook that was exported from NVivo.

Twitter Data

Nodes: Tweets

Name	Description	References
Abuse	Tweets with the term abuse	95
Annoying	Tweets that use “annoying” could be linked to cyberstalking	29
Creep	Tweets that mention creep linked to cyberstalking	150
Fear	Tweets about potential fear of cyberstalking -living in fear of cyberstalking	60
Follower(s)	Tweets about follower or followers or following being compared to cyberstalking	41
Gender	Gender of who the tweet is about	41
Harassment	Tweets that use Harassment	46
Messaging	Tweets with messaging and could be linked with cyberstalking	41
Relationships PP	Tweets about past pr present relationships that can be linked to cyberstalking	44
Scared	Tweets with scared linked to cyberstalking	35
Stalker	Tweets linked to how stalking a can be done on twitter	139
Technology	Use of technology	43
Unwanted	Tweets linked to unwanted and cyberstalking	50
Victim	Tweets associated with victim	40
Violent	Tweets about being violent cyberstalking	21

Nov 11, 2020 1

Figure 18. Finalised NVivo Codebook from the twitter data collected

twitterchart

Unigram	Month	Tweets (5000)	Tweets	Re-Tweets	Hashtags
Abuse	August	95	38	57	65
Annoying	August	29	10	19	15
Creep or Creepy	July	150	88	62	95
Fear	August	60	34	26	33
Follower or Follows	July	41	23	18	27
Gender	July	41	29	12	23
Harassment	July	46	18	28	31
Messaging	July	41	27	14	25
Relationships P/P	May	44	29	15	31
Scared	May	35	16	19	29
Stalker	May	139	96	43	88
Technology	June	43	19	24	26
Unwanted	June	50	27	23	35
Victim	June	40	15	25	22
Violent	May	21	12	9	10

Figure 19. Twitter Dataset: illustrating the unigrams, month, tweets (5000), tweets, retweets, and hashtag

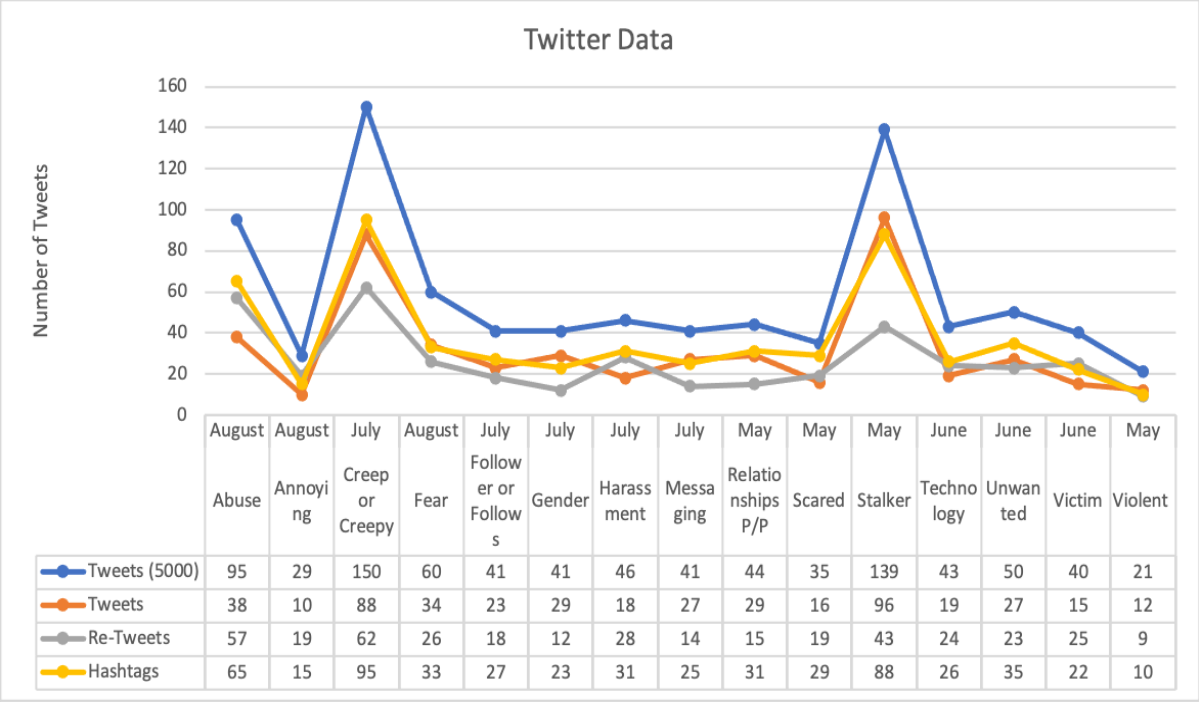
In the above figure, Figure 19, the researcher sifted through 5000 tweets and coded them individually based off the unigram that tweet corresponded too. Subsequently, there was countless tweets to sift through, the researcher wanted to double check that only the tweets that were being used were in fact in correlation to cyberstalking and in English for this study. Once that was finished the researcher was left with this above table that was put into R and graphed and analysed even further. The breakdown of the data and columns in Figure 19, are the fifteen unigrams that were used alongside other information that was used within R. Some of the other information that is presented in the figure above is the month each unigram was searched on twitter and exported from Twitter. The number of tweets that were taken from a thread of 5000, that shown connections regarding cyberstalking. The column

that is next is the tweet column which has the number of tweets, that were original tweets which occurred from the 5000.

Succeeding, is the re-tweet column is the same as the previous column but within concerns to re-tweets. Unlike the column before (tweets) that column is original tweets. The re-tweet column is for tweets that were re-tweeted by many users not tweeted from one person individually. Lastly, the final column is the use of hashtags, out from the total number of tweets that were selected, some of those tweets had hashtags and others did not. Likewise, the researcher thought this part of the data would be interesting to have a further look into, because hashtags are used to pair a tweet into a certain genre. Tweets with hashtags are linked to that hashtag and will be more likely seen from many users on that platform. Whereas a tweet with no hashtag will mostly likely not be as popular or seen by many. Therefore, the researcher wanted to see if there is a correlation with using hashtags or not. Although, it is important to mention, the researcher did not investigate this further for this project but made a note and as thought about proceeding in the future. The researcher thought it was important to have these columns in comparison to others, because as previously stated, the unigrams were not searched under any correlation to cyberstalking nor had any hashtags attached to them while being searched for. As for the month that these unigrams were exported is an interesting, but not overly important. The researcher thought the month could be interesting to have, to see if time of year or season has an effect or influences on how each of the unigrams are being used or tweeted. Again, the researcher noted this and could possibly ensue with this in a future study. In addition, the tweets and re-tweets columns are just as important, if not most attention-grabbing. As you can see, the comparison between the two are split some unigrams have more re-tweets than tweets, while others are the opposite. Thus, making this

research more fascinating, if individuals are tweeting about cyberstalking or their tweet links to cyberstalking. One would think another individual who has experienced the same in the matter, would simple re-tweet the tweet rather than tweeting about it themselves. Which could possibly make this research even more imperative because the many different layers it possesses and all of the possible ways it could be conducted. However, the researcher conducted the study in this manor based off time restrictions and what he or she had access too.

Immediately after the data was collected and coded into nodes within NVivo the data was then exported into excel and graphed. It is remarkable to see how each unigram and its data is compared to each other. That can be shown in the Figure below.



	August Abuse	August Annoying	July Creep or Creepy	August Fear	July Follower or Follows	July Gender	July Harassment	July Messaging	May Relationships P/P	May Scared	May Stalker	June Technology	June Unwanted	June Victim	May Violent
Tweets (5000)	95	29	150	60	41	41	46	41	44	35	139	43	50	40	21
Tweets	38	10	88	34	23	29	18	27	29	16	96	19	27	15	12
Re-Tweets	57	19	62	26	18	12	28	14	15	19	43	24	23	25	9
Hashtags	65	15	95	33	27	23	31	25	31	29	88	26	35	22	10

Figure 20. Line and point chart visualisation of the collected Twitter Data

In the exceeding figure, Figure 20, this figure correlations to the dataset that was imported from Excel. Thus, it is imperious to show case how each variable is acquaintance to one another, but also can contradict one another. Respectively, again

each unigram is broken down into tweets (5000), tweets, re-tweets, hashtags, and the correlation by month the unigrams were searched and exported from Twitter. As you can see from the above figure, the unigrams “creep or creepy” and “stalker” are the two favourable unigrams. For instance, these two unigrams are the most used, but do not have the most hashtags used or the most re-tweets used. The unigram “stalker” does in fact have the most tweets out of the unigrams, but that can be argued merely because of the negative connotation of the unigram itself. Also, to put into factor the two months that these unigrams were captured for this study. July and May are the two months these unigrams were being taken from Twitter with the use of NCapture. For instance, now is there any relationship with those two months and the time of the year with how those two unigrams were or are being used within social media, that is an intriguing concept to investigate or even investigate later.

4.2.2 R Programming

Formerly, the dataset was conducted in Excel after being imported into NVivo with the use of NCapture. The dataset was cleaned up and saved as a csv file and then imported into RStudio and coded. For explanation reasons it is important to know, the researcher is using a MacBook for intent purposes as to explain how R was used and what he or she did within the programme. Therefore, with all the functions used or how the data was saved and how it was handled and how the process is being explained, is because the researcher used his or her MacBook. However, each program can be used on any computer or software programme. The functions that were used in R with the data that was collected is as follows:

- twitter <- read.csv("FDtwitterchart.csv")
- head(twitter)
- colnames(twitter)

        - colnames(twitter) <- c("Unigrams", "Month", "Tweets5000", "Tweets", "ReTweets", "Hashtags")
    - head(twitter)
    - tail(twitter)
    - str(twitter)
    - summary(twitter)

The first three lines of codes are for importing the data, first importing that data from the researcher's computer. Using the code twitter <- read.csv("FDtwitterchart.csv") as previously stated, the dataset is saved as a csv file for the purpose of it being imported into RStudio. Secondly, checking the top of the dataset using: head(twitter), and followed by the column names: colnames(twitter) and making those column names for this dataset colnames(twitter) <- c("Unigrams", "Month", "Tweets5000", "Tweets", "ReTweets", "Hashtags"). The lasty four lines of codes are again the head(twitter) and tail(twitter) which are for the head of the table and the tail of the table. As well as str(twitter) which is used for the structure or debrief of the dataset or data frame and then lastly summary(twitter) is for summary of the dataset in RStudio. Once these codes are run within RStudio the codes are run by pressing command and enter after the line or code or as the line is highlighted.

After the above codes are run in RStudio in the console section this is what is shown and what is being used within RStudio as the dataset.

Shown below is the top part of the summary of the dataset with using head(twitter):

>head(twitter):

	Unigrams	Month	Tweets5000	Tweets	ReTweets	Hashtags
1	Abuse	August	95	38	57	65
2	Annoying	August	29	10	19	15

3	Creep or Creepy	July	150	88	62	95
4	Fear	August	60	34	26	33
5	Follower or Follows	July	41	23	18	27
6	Gender	July	41	29	12	23

Again, while using tail(twitter) that code shows the bottom part of the summary of the dataset, that is displayed like this:

>tail(twitter):

	Unigrams	Month	Tweets5000	Tweets	ReTweets	Hashtags
10	Scared	May	35	16	19	29
11	Stalker	May	139	96	43	88
12	Technology	June	43	19	24	26
13	Unwanted	June	50	27	23	35
14	Victim	June	40	15	25	22
15	Violent	May	21	12	9	10

As previously stated, the str(twitter) function portrays the data frame, however the researcher went and changed that original data frame before graphing and charting the twitter dataset. This step was important and will be explained in further detail, here is how the original data frame would look before the researcher used a few lines of code to change the data frame slight to benefit its progress in RStudio.

```
> str(twitter)
'data.frame': 15 obs. of  6 variables:
 $ Unigrams: chr "Abuse" "Annoying" "Creep or Creepy" "Fear" ...
 $ Month: chr "August" "August" "July" "August" ...
```

```
$ Tweets5000: int 95 29 150 60 41 41 46 41 44 35 ...
$ Tweets: int 38 10 88 34 23 29 18 27 29 16 ...
$ ReTweets: int 57 19 62 26 18 12 28 14 15 19 ...
$ Hashtags: int 65 15 95 33 27 23 31 25 31 29 ...
```

As you can see there are 15 obs. of 6 variables in this data frame. If we take a closer look, you can see how Unigrams is classified as a chr (character) and Month as well is classified as a character then the list of each unigram and month is followed. Furthermore, Tweets (5000), Tweets, ReTweets, and Hashtags, are all classified as int (integer) trailed by their respectable values. Conversely, the data frame needed some alterations. The reasoning as to why the researcher makes these alterations is explained further on with the appropriate codes that are used in similar situations

Using tail(twitter) this code illustrates the last six unigrams and their columns and information. The last code that was listed above is summary(twitter), by means of this code clarifies all of the information that is linked to the dataset that was imported and was used for this study. Summary(twitter) is an important line of code to have a custom for using to gather more insightful information about the dataset.

```
> summary(twitter)

   Unigrams           Month              Tweets5000
   Length:15          Length: 15         Min.: 21.00
   Class: character   Class: character   1st Qu.: 40.50
   Mode: character    Mode: character    Median: 43.00
                                         Mean: 58.33
                                         3rd Qu.: 55.00
                                         Max.:150.00

   Tweets             ReTweets           Hashtags
   Min.:10.00         Min.: 9.00         Min.  :10
   1st Qu.:17.00      1st Qu. :16.50     1st Qu. :24
   Median:27.00       Median :23.00      Median :29
   Mean:32.07         Mean: 26.27        Mean :37
   3rd Qu.:31.50      3rd Qu. :27.00     3rd Qu. :34
   Max.:96.00         Max. :62.00        Max. :95
```

As the twitter dataset is imported into RStudio and the adjustments are made accordingly. The next line of codes is for the graphing and charting portions in RStudio. Although, as mentioned before the str(twitter) function or code is needed, the next list of functions or codes are for slight alterations that are needed to graph or chart this data frame for the twitter dataset.

- factor(twitter$Unigrams)
- twitter$Unigrams <- factor(twitter$Unigrams)

- factor(twitter$Month)
- twitter$Month <- factor(twitter$Month)

- factor(twitter$Hashtags)
- twitter$Hashtags <- factor(twitter$Hashtags)

- summary(twitter)
- str(twitter)

The above list of codes is used to set certain variables as factors which help with the graphing portion in RStudio. As mentioned earlier, Unigrams and Month were classified as characters and Tweets (5000), Tweets, ReTweets, and finally Hashtags were all integers. Moreover, for the purpose of this study these codes are used to change the variable from character or integer to factor, which is the standard application in RStudio. For instance, as you can see above the first six lines of codes does simply that. Factor(twitter$Unigrams) is taking the column unigram in the twitter dataset and making that a factor. Thus, the code is run as twitter$unigrams <- factor(twitter$Unigrams). Moreover, the researcher did the same to the columns Month and Hashtag: factor(twitter$Month), twitter$Month <- factor(twitter$Month), factor(twitter$Hashtags), twitter$Hashtags < -factor(twitter$Hashtags) to make the graphing features have different x and y axes as well as different variables. Finally, summary(twitter) and str(twitter) are run to see if the applicable changes were made

appropriately. Below is the final str(twitter) that shows the suitable modifications that were made.

- str(twitter)

- 'data.frame': 15 obs. of 6 variables:
- $ Unigrams: Factor w/ 15 levels "Abuse","Annoying",..: 1 2 3 4 5 6 7 8 9 10 ...
- $ Month: Factor w/ 4 levels "August","July",..: 1 1 2 1 2 2 2 2 4 4 ...
- $ Tweets5000: int 95 29 150 60 41 41 46 41 44 35 ...
- $ Tweets: int 38 10 88 34 23 29 18 27 29 16 ...
- $ ReTweets: int 57 19 62 26 18 12 28 14 15 19 ...
- $ Hashtags: Factor w/ 14 levels "10","15","22",..: 12 2 14 10 7 4 9 5 9 8 ...

4.2.3 Graphing

Graphing within R Programming is an advanced tool and is the next stage within the research process. RStudio, which is the console where all the coding takes place has many packages and graphing marital that is already within the programme. The researcher used many of the functions within RStudio to his or her advantage. Additionally, in order to make sure R programme itself has a function called ggplot, the individual while in RStudio can run this line library(ggplot2) within the console. After the above line of code is run within the console, then graphing the data can be managed. This line of code seen here is for what would be the x and y axes: ggplot(data=twitter, aes(x= ReTweets, y= Hashtags)). Below is the ggplot that was created within R Programming using the above lines of codes.

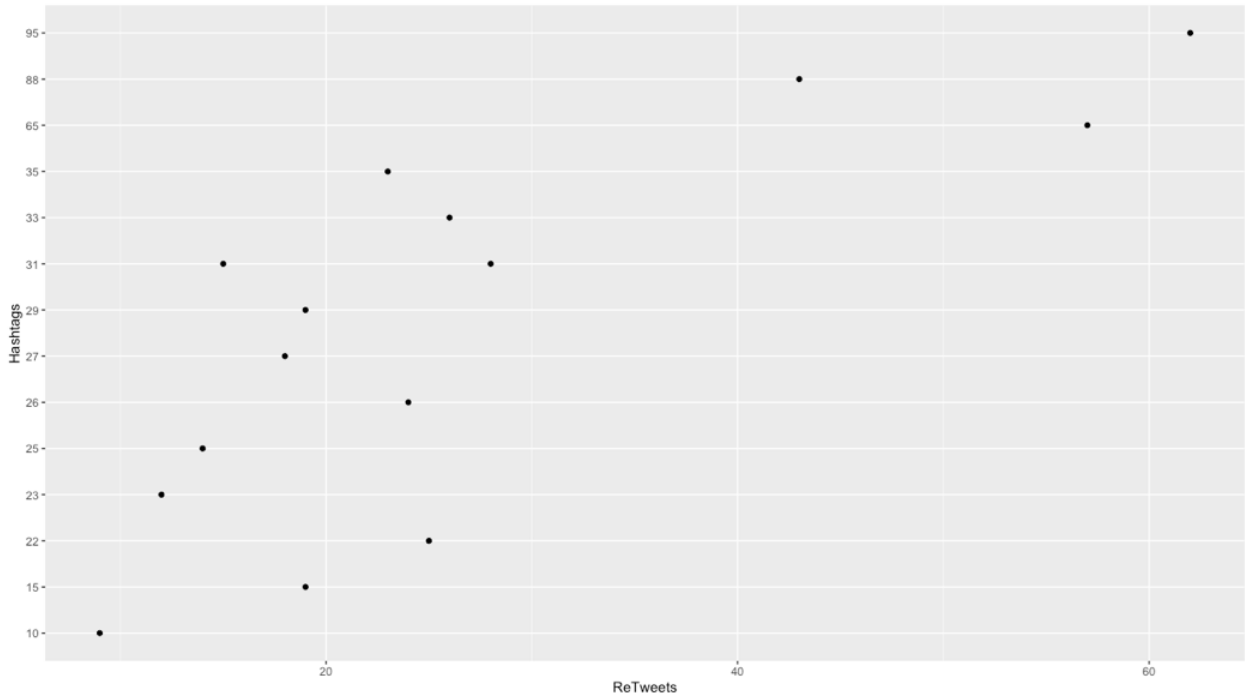


Figure 21. ggplot detailing hashtags v. retweets within the Twitter data

In addition, to the standard graphing layouts within R, the addition of colour, sizes, and many other variables within ggplot will help to showcase the dataset. There are multiple and different strategies or ways the data could be broken-down and displayed. As you can see below colour was added and the size of the points are larger. For reference, the colour indicates the month each tweet was taken from Twitter and the size of the points indicates the tweets out of 5000.

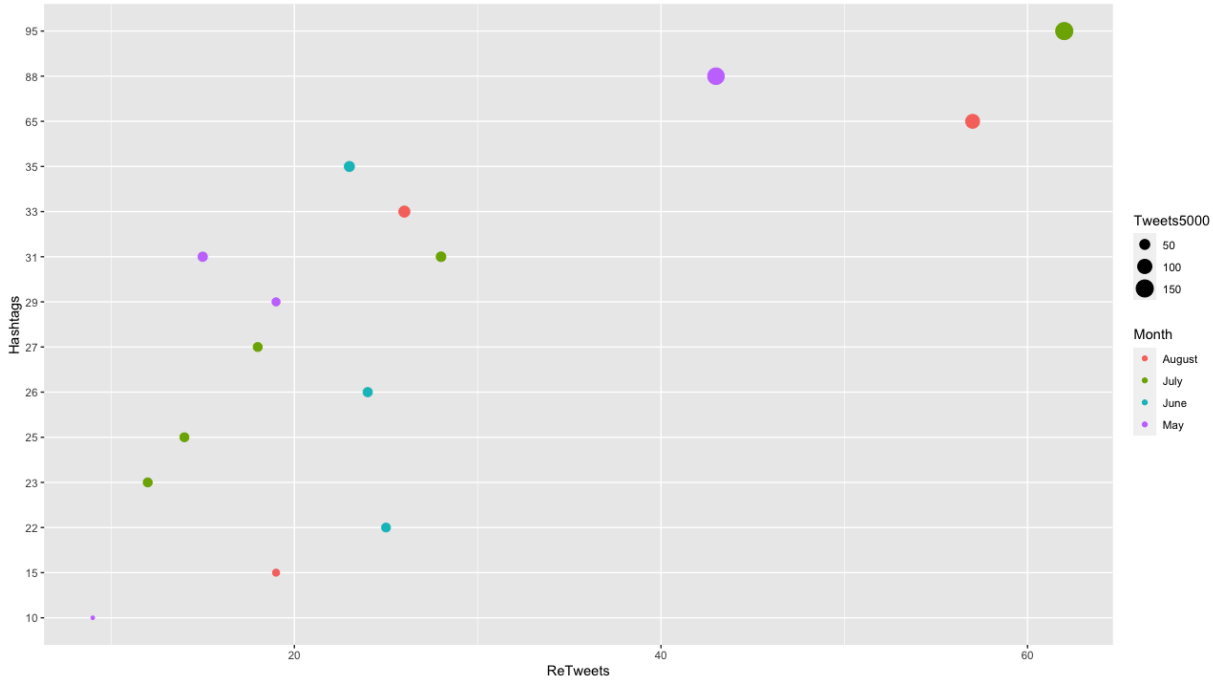


Figure 22. ggplot detailing hashtags v. retweets: colour by month and size of points by tweets (5000) taken from the Twitter data

Showcased above in Figure 22, the axes are the same however, the size of the points on the graph are sized by the Tweet (5000), the more Tweets (5000) the bigger the point. Also, the colour of the points is in correlation to the month, for example: August= red, July= green, June= light blue, May= purple and those colours are given by RStudio the researcher did not picks those colours. Overall, it is interesting to see how the data is spread throughout what month it was collected in comparison to one another. As previously mentioned, does month in fact have any weighting on weather certain or different unigrams are being used more than the others, but that factor was not looked into or investigated further during this study. In Figure 23, the chart is showing a breakdown of each month and how many tweets were coded or saved as a node in NVivo in total as it could be beneficial for further research and findings

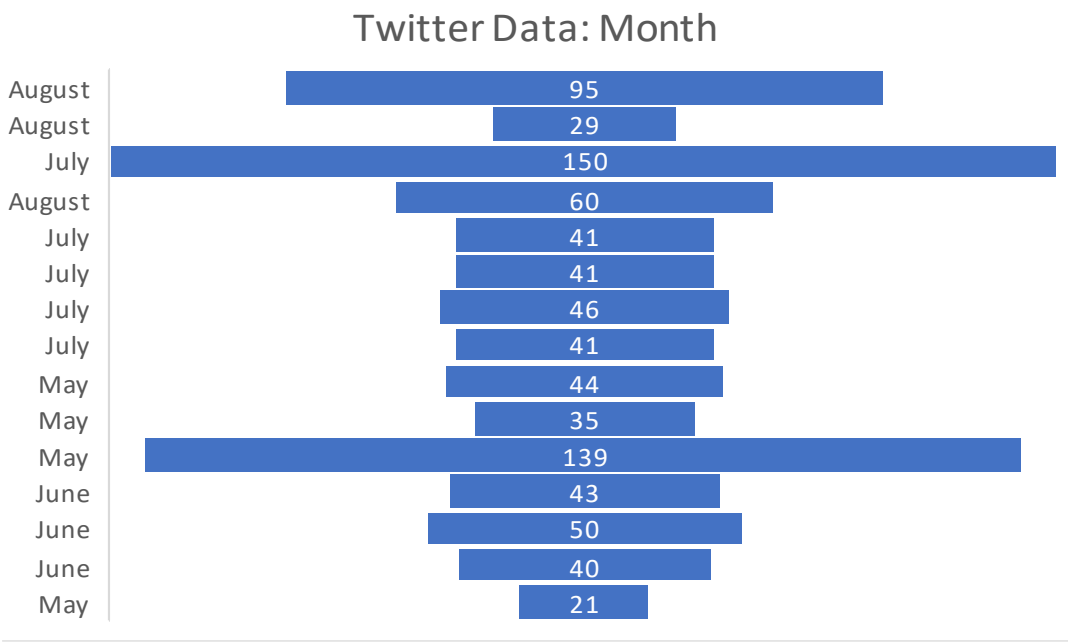


Figure 23. Twitter Data chart overall tweets in correlation by month

As stated earlier in this section, RStudio has many graphing or chart opportunities at hand. Ggplot was already mentioned and shown, but a few other graphs can be highlighted for this study. Some of the graphs are histogram and density charts, mapping v setting, graphing with layers or points, boxplots, scatterplots, adding facets and even much more. Figure 24 is a great example for how relative graphing with layers and points/ lines are; the x axes is Tweets, and the y axes is ReTweets. The colour is again by month and the points are sized by hashtags, the larger the point the more hashtags were used.

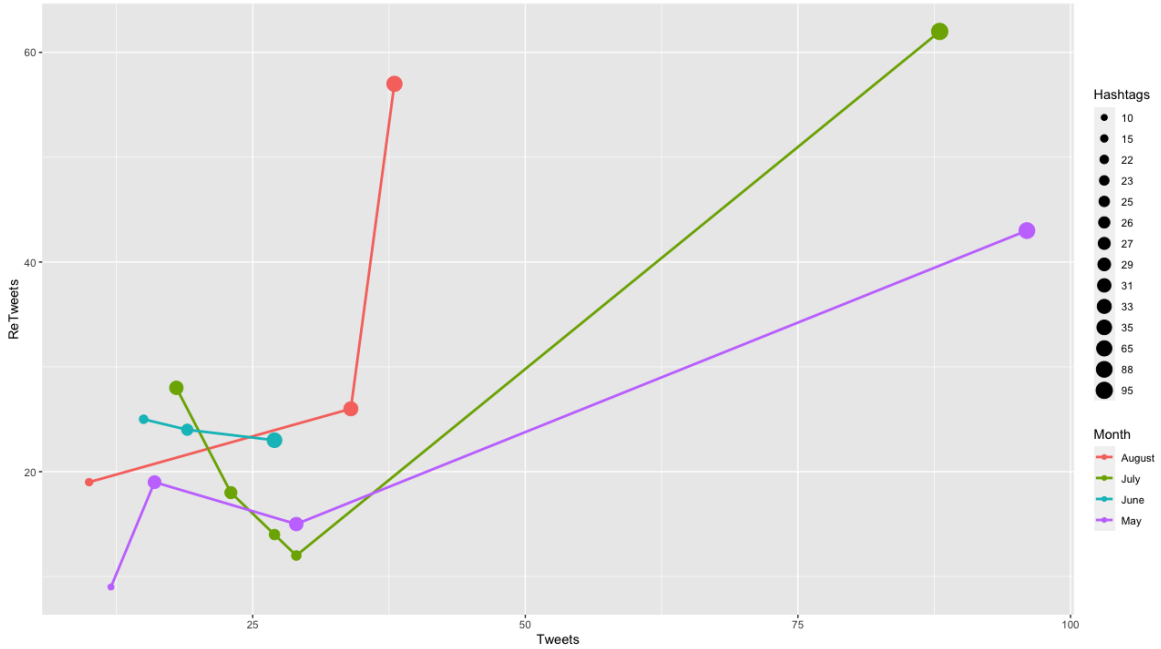


Figure 24. RStudio Layer and line/Point Graph

4.3. Random Sample Dataset in RStudio

The researcher thought it would be beneficial to conduct another data set in R Programming, to test the word frequencies and weighting of terms in comparison to the preliminary data set and the unigrams that was already tested and analysed. The data set that was being used is a very lengthy and large. The researcher only took one part at a time of each of the multiplexed data set and imported it into RStudio.

4.3.1 Dataset and word frequency of preliminary dataset

Before, the researcher mentions the random sample data set. As a reminder the research from the preliminary data set, here are the 15 unigrams that the researcher has chosen within regards to cyberstalking:

- Abuse
- Annoying

- Creep or Creepy
- Fear
- Follower or Follows
- Gender
- Harassment
- Messaging
- Relationships P/P
- Scared
- Stalker
- Technology
- Unwanted
- Victim
- Violent

As well as the number of tweets (5000) each unigram connected with cyberstalking:

Unigram	Tweets (5000)
Abuse	95
Annoying	29
Creep or Creepy	150
Fear	60
Follower or Follows	41
Gender	41
Harassment	46
Messaging	41

Relationships P/P	44
Scared	35
Stalker	139
Technology	43
Unwanted	50
Victim	40
Violent	21

Regarding the information above and looking over each unigram the most common unigram or top choices that would be considered would: stalker, creep or creepy, fear, and maybe even follower / follows. However, the data shows that the unigram creep or creepy has the majority with regards to the 5000 tweets in correspondence to cyberstalking and stalker comes in a close second. The researcher was astounded that the unigrams abuse and unwanted were as elevated in comparison to technology. Since the topic is focused on cyberstalking the researcher thought it would be essential to that the unigram technology would be one of the higher-ranking terms. Nevertheless, when the researcher thought it would be interesting to run a word frequency of all the tweets collected for each unigram together and see if any of the unigrams themselves are a part of the results, the results

were astonishi

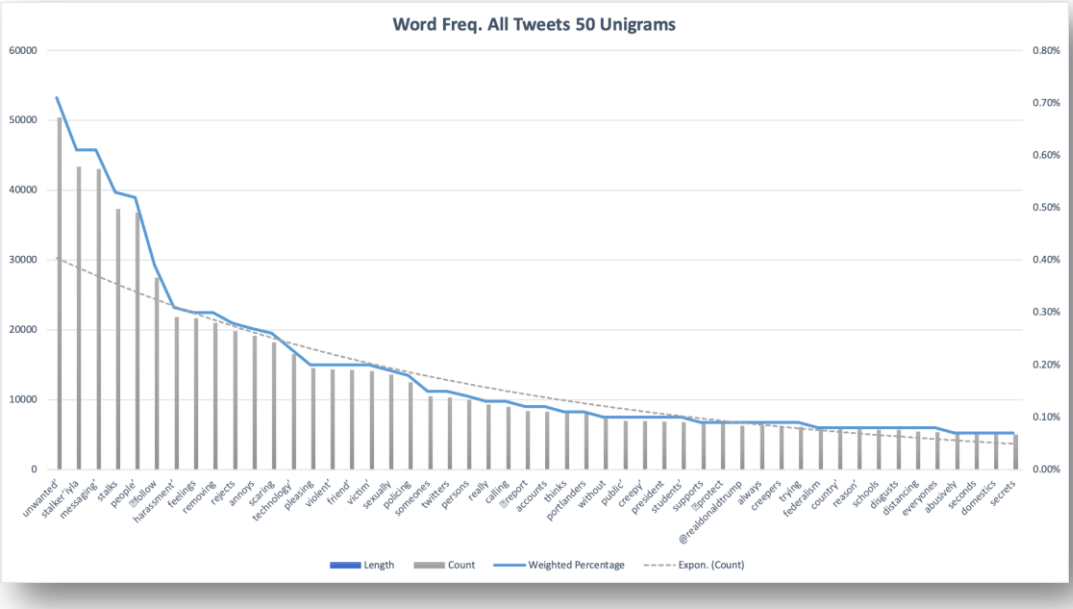


Figure 25. NVivo Word Frequency: taking from the Preliminary Dataset

From the word frequency the above image shown in Figure 25, this chart illustrates 50 of the most used terms or unigrams in all of the Twitter threads that were captured for the preliminary data set. Upon looking closely 11 out of the 15 unigrams are ranked 30 out of the top 50 unigrams. For instance: *unwanted: is 1, stalker: is 2, messaging: is 3, follows: is 5, harassment: is 7, annoys: is 11, scarring: is 12 (*which can be similar to scared*), technology: is 13, violent: is 15, victim: is 17, and lastly creep: 30.* It is important to also point out that the unigrams: creepers, abusively, and stalks are also ranked within this chart. Additionally, the preliminary data set had vast quantity of tweets for each unigram that could have changed the direction of the study depending on the outcome. It is interesting to see the final result, because even though tweets were search by unigram it is fascinating to see how each unigram falls within a word frequency. Even though, the unigrams are linked to the tweets themselves, the

unigram are not linked to any advancement towards cyberstalking and yet 11 out of the 15 unigrams are in the ranked in the top 30 most frequent terms. However, with that being said that does not mean the majority of tweets those unigrams have are in favour to cyberstalking. As mentioned, before it is important to remember that these unigrams that were selected were done off of a test run. The researcher searched on twitter before the data collection process was done, to gather an understanding of how the term cyberstalking was perceived on Twitter. The researcher searched "cyberstalking" on twitter to see what information or what the correspondence would have been. Continuously, that search thread was put into NCapture and then uploaded into NVivo and run through a word frequency with no alterations to that dataset. Although, the term cyberstalking was indeed searched there was no advanced search attached; therefore, not all the tweets would be correlated to cyberstalking, some tweets are random in evaluation. However, the word frequency that was conducted helped illustrate how this research can be directed and gave an insight on which unigrams or terms to emphasis on for the overall purpose for this research.

4.3.2 Importing the random sample of tweets: Dataset(s) 1-5 in RStudio

As stated in the introduction to this section, the researcher thought it would be beneficial to conduct another data set within R Programme. The data set that was used is extremely extensive and considerable. Therefore, the researcher imported the data set into RStudio in five different csv files. The next sections or segment of this chapter will pay close attention to data set(s) 1 through 5, each file had a total of 50,000 (1,500,000 data points in total from the dataset) tweets, once uploaded into R the researcher then cleaned the dataset which will be highlighted in more detail below. The figures that are shown below are in correlation with these datasets and the preliminary data set that was mentioned and used for chapter three. The results from

the preliminary data set and the random sample data set, will inform the researcher if there is a correlation with the unigrams along with the tweets from the random sample dataset in comparison to cyberstalking. These five datasets were preferred to be the prime centre data sets in this section, because all of the data set's support and illustrate a correlation between the unigrams and tweets attached to them. However, only two data sets will be shown in this chapter and the other three will be in the appendix section at the end of the thesis. In addition, all the remaining figures and tables for the data sets will be explain in detail further on as this chapter continues.

Moreover, once the data sets were imported into RStudio the data cleaning process could be underway and then completed. The researcher then ran a word frequency on the dataset, remember all of these steps were done to each random sample data set. After the word frequency was concluded the researcher took the terms and the count for each data set and ranked them in excel to have a better look in comparison to the bar-plot that was done inside RStudio.

Furthermore, before the bar-plot and table charts are looked at in further detail, here is how the data set was established before being imported into RStudio. Below is data set 1, which has 50,000 random sample of tweets that were imported into RStudio and coded for a better understanding for this study. However, before each data set was imported into RStudio, the researcher had to add headings to each csv data set file(s). This was done to help with the text mining process and to write the functions and strings in R with a bit more ease. For example, each csv file (e.g., Data 1.csv, Data 2.csv and so on) contains the 50,000 tweets as stated previously, with the following 6 fields:

- target: the polarity of the tweet (0 = negative, 2 = neutral, 4 = positive)
- ids: The id of the tweet

- date: the date of the tweet
- flag: The query. If there is no query, then this value is NO_QUERY.
- user: the user that tweeted
- lastly, text: the text of the tweet

The researcher opened Excel, and made an extra row and added these headers for the text mining process within RStudio the headings are: n/a, ID, Date, FALSE, Username, and text:

- n/a = which was for the target polarity of the tweets:(0, 2, 4 in the dataset)
- ID= for the tweet ids
- Date= for the date of the tweet
- FALSE= for the query, there was no value so in the columns NO_QUERY
- Username= for the user that tweeted
- And lastly, text= for the tweet itself

After the researcher made the slight change and added the headers to each column and updated/replace the saved csv file on his or her desktop. The csv file is ready to be imported into RStudio. Now importing data or files into RStudio can be done in many different ways. The researcher imported csv files from his or her desktop into R. However, before that is done the individual who is using the software must know where or what their working directory is or set up as. For example, the working directory can be the individual user who is the owner of the computer that the software was downloaded on, or the working directory can be set manually. To set the working directory manually, that can be done by; selecting it in RStudio click “files” then click the file you want to select as your working directory. Once that is selected, click “more” which is on the toolbar that is underneath files, plots, and packages and so on.

Immediately, after more is clicked or selected a drop-down bar will appear and then finally select the option: set as working directory.

Nonetheless, the string function to get the whereabouts of the working directory so the individual who is using RStudio knows what his or her directory is as follows: getwd(). When the string is run shown in the console would be the working directory. Thus, the above information can be used to either change the working directory or keep it the same. The next few lines of code will be what the researcher has done to import the data, the csv file from the working directory that he or she has set up as from their desktop. It is important to know that the researcher did these steps for each data set. Formerly, to import any csv file data set it is imperative to use the string: read.csv() function.

I. tweets <- read.csv(file="~/Desktop/Data5.csv", header= T)

    a. this is how the researcher imported the data set. “tweets” was used as the name for the data set. The function read.csv() is used for csv files only, in the () the function: file=, is used for RStudio to know where the file is saved, but is only used when the csv file is saved to the desktop not the users working directory folder. It is essential to note that the file and where it is saved is put inside “” (quotations) or the file will not be imported. Lastly, the function header= TRUE or FALSE; is used, because as previously mentioned the researcher made headers for each dataset csv file. Therefore, in this case headers are true and was stated in RStudio so the programme new which column falls under what header. However, if they were false and not needed then use FALSE. Reminder, the words true and false need to be capitalised within RStudio for the function to work.

II. str(tweets)

a. the function str(), is used to look at the overall data frame for example, this is what the str(tweet) function output in the console:

'data.frame': 50000 obs. of 6 variables:

$ n.a : int 0 0 0 0 0 0 0 0 0 0 ...

$ ID : int 1971570900 1971571167 1971571296 1971571313 1971571456 1971571463 1971571597 1971571647 1971571833 1971571835 ...

$ date : chr "Sat May 30 07:24:19 PDT 2009" "Sat May 30 07:24:21 PDT 2009" "Sat May 30 07:24:22 PDT 2009" "Sat May 30 07:24:22 PDT 2009" ...

$ FALSE. : chr "NO_QUERY" "NO_QUERY" "NO_QUERY" "NO_QUERY" ...

$ username: chr "Gemma_Rigby" "chaubaole1507" "CeriseJC" "TheSwelleLife" ...

$ text : chr "Can't believe I have to wait another 6 months for my phone contract to end! I'm bored now!!! The 12 month contr"| __truncated__ "When did I felt so lonely? " "ugh. a huge headache, coughing constantly, legs feeling week, and feeling like throwing up. This sucks beyond compare " "Got to go clean now, knowing it will be messed up again by tomorrow. " ...

III. If we take a closer look at the data frame the headers that were made to the csv file in the Excel follow the $ symbol. All of the data set is broken up into 6 variables and 50000 observations. Numeric variables are classified as an integer, any words/lettering is classified as a character. This is the standard setting when a dataset is imported into RStudio. However, to clean this dataset and build a corpus, as well as make a term document matrix and finally plotting the frequent terms in the tweets; the $text vector needs to be changed from character to factor.

a. the function is quite simple: take the dataset which is tweets and the column we need $text make that the start of the function. Followed by the function as.factor which will change $text from character to factor. The string and function to change character to factor is as follows:
    i. tweets$text <- as.factor(tweets$text

4.3.3 Building Corpus and Cleaning the Dataset in RStudio

On the other hand, the reasoning as to why building a corpus is such an important step is because if the dataset is made into a corpus the cleaning, plotting, and word frequency cannot be done. Formerly, following the above steps: importing data set the strings and functions are preformed then building a corpus is the next step of action necessary. Building a corpus is only two lines of code in this situation based off of the data set at hand. Following these strings and functions for building a corpus using the tm packaging within R is as follows:

- library(tm)
- corpus <- iconv(tweets$text, to = "utf-8-mac")
    - research and common practice in R shows: UTF-8 is the most widely used way to represent Unicode text in web pages, and should always use UTF-8 when creating web pages and databases. But, in principle, UTF-8 is only one of the possible ways of encoding Unicode characters.
    - for instance, the utf-8-mac is the utf 8 version of a text after application of Unicode normalization NFD (e.g., accented characters are represented by the base character plus a combining accent character), with certain codepoint ranges excluded from the decomposition operation
- corpus <- Corpus(VectorSource(corpus))

- "Corpus" is a collection of text documents within RStudio. VCorpus in tm refers to "Volatile" corpus which means that the corpus is stored in memory and would be destroyed when the R object containing it is destroyed. In order to create a VCorpus using tm, it is needed to pass a "Source" object as a parameter to the VCorpus method. The sources available using this method: getSources()
- For reference:
    - input <- c('This is line one.',' And this is the second one')
    - Create the source: vecSource <- VectorSource(input)
    - Then create the corpus: VCorpus(vecSource)

- inspect(corpus[1:5])

The last line inspect(corpus[1:5]), inspects the first five lines of the data set column tweets. Therefore, it will show the first five tweets in the dataset. Next, the dataset needs to be cleaned, which can be a length process. These functions below show how to clean the data set.

Beneath are the examples of code, the first line of code is taking the data set and making it lowercase, using the tolower function and then inspecting the first five tweets again. However, the function: tm_map, is used to apply transformation functions to the corpus. After each line of code, accordingly the first five tweets are inspected to make sure the string and functions were done properly.

- corpus <- tm_map(corpus, tolower)
- inspect(corpus[1:5])

Moreover, the removal of the punctuation using the removePunctuation function is vital for the text mining within this research. Again, the tm_map function is used and will be used throughout the remaining stings and functions. As well as, the removal of

numbers, which is done because numbers are not beneficial for this research. Therefore, the removeNumbers function is used within the corpus and tm_map.

- corpus <- tm_map(corpus, removePunctuation)
- inspect(corpus[1:5])
- corpus <- tm_map(corpus, removeNumbers)
- inspect(corpus[1:5])

Lastly, the final steps for cleaning a dataset are as follows. The elimination of stop words is imperious because they are the minimal every word(s) that are used with writing or expression on twitter. For example, Figure 26 identifies the stop words that R eliminates:

```
> stopwords("english")
  [1] "i"          "me"         "my"         "myself"     "we"         "our"
  [7] "ours"       "ourselves"  "you"        "your"       "yours"      "yourself"
 [13] "yourselves" "he"         "him"        "his"        "himself"    "she"
 [19] "her"        "hers"       "herself"    "it"         "its"        "itself"
 [25] "they"       "them"       "their"      "theirs"     "themselves" "what"
 [31] "which"      "who"        "whom"       "this"       "that"       "these"
 [37] "those"      "am"         "is"         "are"        "was"        "were"
 [43] "be"         "been"       "being"      "have"       "has"        "had"
 [49] "having"     "do"         "does"       "did"        "doing"      "would"
 [55] "should"     "could"      "ought"      "i'm"        "you're"     "he's"
 [61] "she's"      "it's"       "we're"      "they're"    "i've"       "you've"
 [67] "we've"      "they've"    "i'd"        "you'd"      "he'd"       "she'd"
 [73] "we'd"       "they'd"     "i'll"       "you'll"     "he'll"      "she'll"
 [79] "we'll"      "they'll"    "isn't"      "aren't"     "wasn't"     "weren't"
 [85] "hasn't"     "haven't"    "hadn't"     "doesn't"    "don't"      "didn't"
 [91] "won't"      "wouldn't"   "shan't"     "shouldn't"  "can't"      "cannot"
 [97] "couldn't"   "mustn't"    "let's"      "that's"     "who's"      "what's"
[103] "here's"     "there's"    "when's"     "where's"    "why's"      "how's"
[109] "a"          "an"         "the"        "and"        "but"        "if"
[115] "or"         "because"    "as"         "until"      "while"      "of"
[121] "at"         "by"         "for"        "with"       "about"      "against"
[127] "between"    "into"       "through"    "during"     "before"     "after"
[133] "above"      "below"      "to"         "from"       "up"         "down"
[139] "in"         "out"        "on"         "off"        "over"       "under"
[145] "again"      "further"    "then"       "once"       "here"       "there"
```

[151]	"when"	"where"	"why"	"how"	"all"	"any"
[157]	"both"	"each"	"few"	"more"	"most"	"other"
[163]	"some"	"such"	"no"	"nor"	"not"	"only"
[169]	"own"	"same"	"so"	"than"	"too"	"very"

Figure 26. Showcasing the elimination of these stop words in R Programme

Consequently, the strings and functions used to eliminate stop words in R are like this:

- cleanset <- tm_map(corpus, removeWords, stopwords("english"))
- inspect(cleanset[1:5])

In addition, removing the URLs are just as important, links to other sources within tweets can be very beneficial. However, for this research and its purpose they are not needed therefore were removed and to do that here are the lines of codes using these string and functions:

- removeURL <- function(x) gsub('http[[:alnum:]]*', '', x)
- cleanset <- tm_map(cleanset, content_transformer(removeURL))
- inspect(cleanset[1:5])

there is another valuable function to use when cleaning up datasets and it is the stripWhitespace function. Which can be shown below, this function is used because up to this point your dataset is missing values such as: stop words, numbers, punctuations. URL(s) etc. Hence, it is important to strip the whitespace within the text to make the tweets more readable/codable.

- cleanset <- tm_map(cleanset, stripWhitespace)
- inspect(cleanset[1:5])

As you can see from beneath the tweets look like normal sentences, however the stop words, white space, numeric value, or URL links were eliminated as mentioned before. When this step is conducted the tweets are no longer coherent sentences, but the information that was removed had to be done for this study. Thus, the below function is run this output is seen on the console:

<<SimpleCorpus>>

- Metadata: corpus specific: 1, document level (indexed): 0
- Content: documents: 5
- [1] cant believe wait another months phone contract end im bored now month contract run
- [2] felt lonely
- [3] ugh huge headache coughing constantly legs feeling week feeling like throwing sucks beyond compare
- [4] got go clean now knowing will messed tomorrow
- [5] still hoping google take world algebra revision

4.4.4 Making a Term Document Matrix

A term document matrix constructs or coerces to a term-document matrix or also known as a document-term matrix. Plus, a corpus is a valuable piece of making a term document matrix. For instance, a corpus is the constructors of either a term-document matrix, document-term matrix, a simple triple matrix (package slam), or a term frequency vector for the coercing functions. That is why making the corpus into a term document matrix helps with graphing and plotting the word frequencies for each data set. The strings and functions used to do this are as follows:

First off, the vector needs to be named in this case it is named tdm, for term document matrix. Then the function TermDocumentMatrix() is used, in the () cleanset, which is the cleaned corpus that was created previously. Followed by control = to set the list of mini word length, which is set to 5, infinitive then closed off to end the string. The second line of code is when the removal of sparse terms is added. The reasoning as to why sparse terms need to be removed from a document-term matrix or term-document matrix is: the numeric for the maximal allowed sparsity in the range from

bigger zero to smaller one. A term-document matrix where those terms from x are removed which have at least a sparse percentage of empty (i.e., terms occurring 0 times in a document) elements. For example, the resulting matrix contains only terms with a sparse factor of less than sparse. In this study the spares had to be changed and set to 0.98 for more of a pure result, the spare ordinally was 100 which would not have given the best results for this study. Lastly, set the vector t, and make it into a matrix.

- tdm <- TermDocumentMatrix(cleanset,
    - control = list(minWordLength=c(5,Inf)))
- t <- removeSparseTerms(tdm, sparse =0.98)
- m <- as.matrix(t)

4.3.5 Plotting Frequent Terms

Plotting frequent terms and obtaining the word frequency for these data sets is a vital and remarkable part of this research. The researcher first made a bar-plot of the most frequent terms in each data set. The bar-plot was made from the row, which are the sums of the vector m. Followed by the most frequently used terms set to 25, so any word that has been used or seen 25 or more time within the tweets was flagged. Lastly, the illustration of the bar-plot, freq, followed by las=2, which is the standard setting it the measurements of the labels on the bar-plot. Finally, the colour of the pot itself which the researcher set to rainbow for a better visual effect. Here are the functions and strings used followed by the bar-plot itself:

- freq <- rowSums(m)
- freq <- subset(freq, freq>=25)
- barplot(freq, las=2, col = rainbow(25))

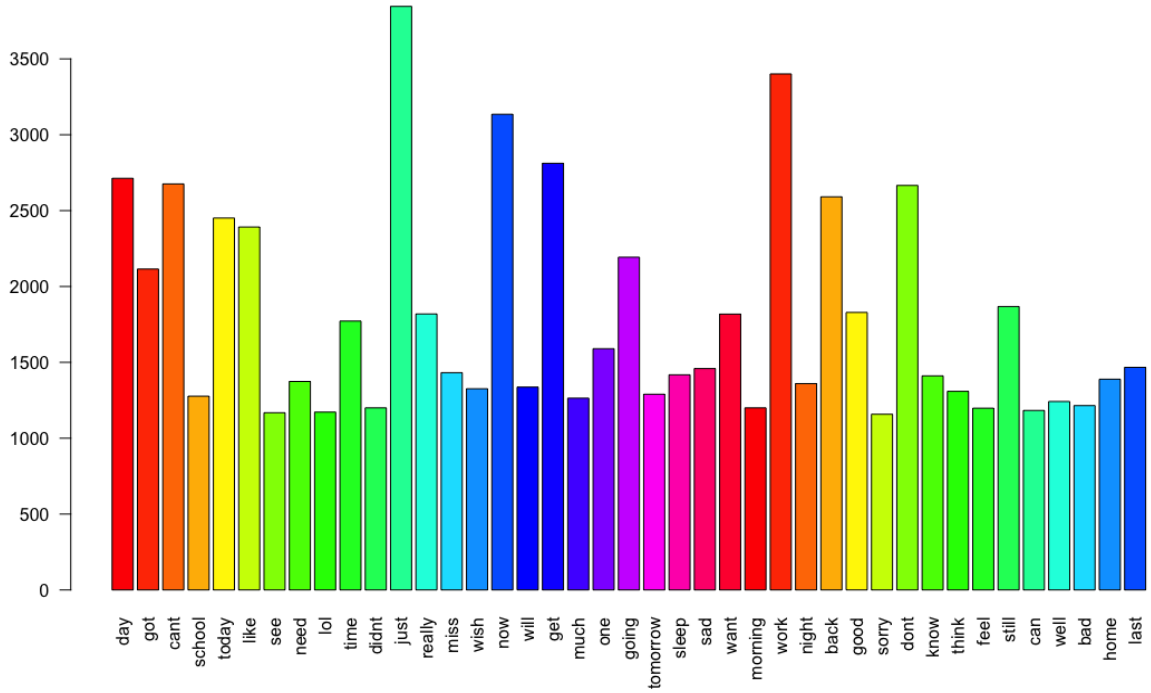


Figure 27. Bar-plot for Dataset 1: taken from the Random Sample Dataset

Secondly, once that the bar-plot is formed and completed the researcher then wanted the exact numeric word frequency for each term. Yes, that is shown in the bar-plot above in Figure 27, however, the researcher wanted a purer observation. Nonetheless, the researcher used the strings and functions below then made a table ranking each term to illustrate the findings in a sharper view. The functions and stings are shown here as well as the table ranking each term.

- freq <- sort(freq, decreasing = TRUE)
- head(freq)
- tail(freq)
- str(freq)
- freq

Freq Word	Count	Rank
just	3846	1
work	3401	2
now	3135	3
get	2812	4
day	2713	5
cant	2676	6
don’t	2666	7
back	2591	8
today	2451	9
like	2392	10
going	2193	11
got	2155	12
still	1868	13
good	1829	14
really	1819	15
want	1818	16
time	1772	17
one	1590	18
last	1467	19
sad	1459	20
miss	1432	21
sleep	1418	22
know	1411	23
home	1389	24
need	1374	25
night	1360	26
wil	1337	27
wish	1326	28
think	1309	29
tomorrow	1290	30
school	1277	31
much	1264	32
well	1242	33
bad	1215	34
didn’t	1200	35
morning	1200	35
feel	1198	37
cant	1183	38
lol	1172	39
see	1168	40
sorry	1158	41

Figure 28. Ranking of Frequent Words: from Dataset 1 within the Random Sample Dataset

4.4. Results and Discussion

Throughout this chapter, all the data sets that have been shown and displayed how they were collected, imported, exported, graphed, and even run within the programmes. However, in this section the researcher will explain the twitter data (random sample data set) in detail and to compare and/ or contrast both the preliminary data set and the twitter data sets with one another. The researcher will use more graphs, charts, and advance searched in R Programme provided from each of the twitter data set to explain how the data along with the preliminary data set is analysed even further. As well as the potential of using different aspects of the data set would or can alter the findings or results, or if those alterations can be beneficial for future studies on this matter.

As mentioned in previous chapters and earlier in this one, RStudio was used to analysis the data in connection to NVivo. Both Programmes were used as an advantage this project or study. Respectively, each programme brought significant findings and remarkable insight to this topic at hand. The researcher noticed while using RStudio with the five datasets of random samples of tweets, that certain terms show similarity to the preliminary data set which was the focus of chapter three. For instance, if we look at the results, we can notice a few terms from the random sample data set that might be correlated to cyberstalking.

The researcher was intrigued to find while looking through the five random sample data set findings. Certain terms that have been seen before in collection of the preliminary data set. For instance, looking at dataset 1s findings and dataset 3s, each bar-plot and ranking have certain words that were seen before in the preliminary data. Note to mention, that the five data sets that were used are random everyday tweets

that have no link or advancement to cyberstalking. With that in mind, it makes the results more significant and incredibly fascinating that there is a similarity to both data sets. For example, this ranking chart beneath was from dataset 3. Some of it was deleted for the illustration purpose, the terms highlighted in yellow have been seen before and are highlighted yellow in table 1 in the appendix which is a word frequency taken from NVivo on the preliminary data set. Furthermore, those reoccurring terms are more emotional terms that keep being represented for example: like, hate, angry, feel, bad, or sad. Nonetheless, the word frequency that was obstructed from the random sample data set, was non-connection terms or themes in comparison to direct cyberstalking. However, that was to be expected since data set that was being used is a random sample data set. It is in fact a vast number of random tweets which none had affiliation to any topic at hand. Therefore, the word frequency that was obtain would have no direct correlation towards cyberstalking. Amongst, seen through each data sets, bar-plots, and tables which are in the appendix there is a similarity to emotional terms that can be used within or as an exaptation to cyberstalking. Also, the words highlighted in orange are interesting terms that have been shown throughout each stage of the data collection and analysis process, these terms might be useful for further looking into with any relationship to cyberstalking.

Word Freq	Count	Rank
really	1879	15
time	1718	17
miss	1759	16
want	1691	18
last	1691	18
home	1576	20
one	1542	21
know	1487	22
sad	1470	23
will	1453	24
night	1373	25

feel	1369	26
think	1365	27
need	1344	28
bad	1338	29
well	1286	31
wish	1255	33
can	1240	34
didn't	1235	35
sorry	1234	36
morning	1202	37
hate	1048	41

Figure 29. Edited: Word Freq, Count, and Rank, Table from Dataset 3, used from the Random Sample Dataset

As the result to continue to show a link in association towards the preliminary data set. Again, looking at the edited table from dataset 2, the same terms that have been seen in both numerous tables in the preliminary data set, these terms seem to be fourth coming in this data set as well. These results slightly impact the research because they can be seen as reoccurring emotional key themes. However, as the researcher is conducting this project, the idea of running an advanced search using each unigram(s) brought forward from the preliminary data set searched within each random sample data set. This would be more advantageous for this study and will show if there is in fact a correlation with the unigrams used and the random sample data set regarding detecting cyberstalking. The outcome or result of the search within RStudio will continue and further the narrative that surrounds the focus point of this research.

Freq Word	**Count**	**Rank**
day	3052	3
work	2893	4
like	2537	8
miss	1756	18
sad	1695	19
sleep	1385	27
didn't	1362	28

night	1353	29
feel	1338	30
think	1332	31
lol	1328	32
tomorrow	1319	33
bad	1312	34
well	1268	35
see	1254	36
sorry	1253	37
hate	1104	40
love	1052	41

Figure 30. Figure 31. Edited: Word Freq, Count, and Rank, Table from Dataset 2, used from the Random Sample Dataset

As you can see, in Figure 30, is another example to the emotional linking or connected between both data sets, the preliminary data set and the random sample data set. The reoccurring themes or terms that are being brought to the researcher's attention and represented are: really, think, feel, need, like, sad, hate, love. Straightaway as previously mentioned, these reoccurring terms have been seen before in the preliminary data set. Is there a correlation between these terms and the vast quantity of tweets they are in with cyberstalking? Furthermore, are these terms linking terms that hint cyberstalking might be taking place for instance, let's use the term hate. If the researcher looks further into the tweets that have mention or flagged up with the term hate, are any of those suggesting or a product of cyberstalking. This can be done with any of the terms that have been seen reoccurring within the data sets being used for future research on this study. Likewise, as stated above, the terms in orange can be investigated much later for future reference.

4.4.1 Advanced Search in RStudio

As the researcher continued with his or her data analysis on the random sample tweets dataset. As previously mentioned, while the results of the frequent words within each data set was not surprising and was to be expected, because the data set was a random sample of tweets correlating to anything not a specific topic. Furthermore,

the reasoning as to why the researcher thought an advanced search of each unigram on each dataset would be very constructive for this study. The data set that will be the prime focus is dataset 1. All the other data sets i.e.: 2-5 are in the appendix section with the same exact breakdown as shown in dataset 1. The researcher searched each unigram individually in each data set within RStudio. The researcher used the grep() function to do the advanced search. The grep() function is used to search a file or text document for anything specific can be a word like in this study or can also be numeric. However, this function does not tell you how many times the word appears, but it does bring forward all the texts that the searched unigrams or word appears in.

In other words, in this case each tweet that has the unigram that was being searched was brought forward to the researcher's attention. Now, as you can imagine there were vast number of tweets for each unigram, because it pulled every tweet that had mentioned the unigram to light. Once, the researcher conducted the grep() function and was given all of the results for each tweet. The grep() function is used as follows: grep("insert word you like to search", followed by the dataset or frame that is being used and the location of the text file: for example tweets$text). The entire function looks like: grep("annoying", tweets$text). Once, the function was run and the console gave the results for each tweet. The following step was taken, the string tweets$text was used again followed by the square brackets [] with the number that was given for each tweet. Consequently, that string would look like: tweets$text[46986].

As seen in the example below, which helps illuminate how the researcher conducted the advanced search within RStudio. The example that is being shown is from csv file dataset 1. Out of the 15 unigrams 4 of the unigrams had tweets that correlate with cyberstalking. Those unigrams are annoying, creep/creepy,

follow/follows, and stalker. Therefore, the unigrams: abuse, fear, gender, harassment, messaging, relationships, scared, technology, unwanted, victim, and lastly violent had no tweets in this csv file that had any correlation to cyberstalking. If we look at the first unigram annoying, the number 46986 which was the number correlated with the unigram used in that tweet. Although, there is only one number shown for annoying there were many numbers as the result of using the grep() function for each unigram to verify each tweet. Hence, the tweet and number that were selected are the only numbers shown below. As you can see, there are sections of tweets in all the data sets that have are highlighted black. Those specific sections have either: an individual's name, twitter username or tweet handle, this information is blocked out because it is to keep the anonymous value to the research. Beneath are the dataset 1 advanced search findings:

Dataset 1.csv (11 Tweets)

<u>Annoying 1</u>

> grep("annoying", tweets$text)

[46968]

> tweets$text [46968]

[1] [redacted] hey [redacted]! i see [redacted] won't stop commenting/stalking you/being annoying. i feel for you, i really do (stalker)

<u>Creep/Creepy 3</u>

> grep("creepy", tweets$text)

[20660 33099 29981]

> tweets$text [20660]

[1] [redacted] actually, honestly that would be creepy if someone I'd just met sent me that

> tweets$text [33099]

[1] 12:20. Online at school. My classmate is a creepy fucking stalker, I hate him (stalker)

> tweets$text[29981]

[1] WHAT THE FUCKKKK, why r random people all the sudden following me, im kinda creeped out

Follow/Follows 1

> grep("follow", tweets$text)

[14454]

> tweets$text[14454]

[1] [redacted] LOL i know what you meant. btw random people follow you....

Stalker: 6

> grep("stalker", tweets$text)

[908 5352 9255 11307 11737 15477 16412 19472 20414 33099 34342 34354 41621 49625]

> tweets$text [9255]

[1] All my places of solitude on line are being taken over by a stalker Is nothing in my world to be sacred anymore?!

> tweets$text [16412]

[1] [redacted]!! another stalker??

> tweets$text [20414]

[1] [redacted] I agree w/ [redacted] Girl u tryna get snatched up? u can have a lowkey stalker watching ur every move right now! don't need that

> tweets$text [33099]

[1] 12:20. Online at school. My classmate is a creepy fucking stalker, I hate him

> tweets$text [41621]

[1] [redacted] my head is pounding and I got a robot stalker... Hope ur day is better

> tweets$text [46968] The unigram stalking was searched to show case this in R

[1] [redacted] hey [redacted]! i see [redacted] won't stop commenting/stalking you/being annoying. i feel for you, i really do

Analysing the results to dataset 1 from the advanced search handled within RStudio. It is pivotal to notice that the unigrams that were picked from the preliminary data are in fact are useful and can potentially detect cyberstalking tendences on Twitter. Within each data set, it is remarkable to see that some of the same unigrams keep flagging tweets associated towards cyberstalking. However, some of the unigrams do not have associated tweets concerning cyberstalking nor even in relation with cyberstalking. Moreover, the results are attention-grabbing, in the sense of within all five datasets the unigram stalker has tweets that appears thus being flagged towards cyberstalking. Also, the unigram follow/follows and creep/creepy have tweets that are the focus in four out of the five datasets. In addition, the unigrams abuse and annoying both show tweets only in two data sets. Lastly, with the unigrams scared and technology only having one tweet flagged in one dataset. On the other hand, the

unigrams: fear, gender, harassment, relationships, messaging, unwanted, victim, and violent. All had zero tweets that were brought to the researcher's attention that foreshadowed cyberstalking in anyway. With, those results could possibly be the outcome of the data set being a random sample of tweets. However, it could possibility be that those unigrams themselves do not showcase a high volume of tweets in relations to cyberstalking. Likewise, that is not implying that these unigrams have zero correspondence but could potentially have a weaker quantity in comparison to the other unigrams that are profoundly reoccurring.

The development of this strategy is imperial to this study, because there is a link between both data sets that are being used by the researcher. Remember, the preliminary dataset was established around cyberstalking and the five random sample of data sets were exactly that a random sample of tweets. Again, mentioned before, the link between the two is imperative, however, if these terms/unigrams and the tweets that are mentioned in the each of the random sample data sets prove to suggest or are a product of cyberstalking, then the correlation is true, and the research done thus far is advantageous.

4.5. Summary

In conclusion to summarise, this chapter describes the research methodology used to analyse the data that was collected and required to address the research questions and to test the hypothesised relationships developed in this study. The chapter begins with a discussion of the data importation and clean-up followed by the programmes in which data was analysed and the approach to using the random sample data set towards the comparison to the preliminary data set and cyberstalking. The chapter then continues with descriptions of how the random sample data set was analysed, the data measurement, as well as the reoccurring terms that are being seen

within all the data sets. Lastly, the advanced search of each unigram that was conducted within RStudio inside each random sample data set.

As the results demonstrate that there can be a correspondence between both data sets and cyberstalking. Henceforth, the researcher thought it would be more valuable to look into all the unigrams and the certain tweets that are flagged up with more detail and to see if the tweets brought to attention are associated to cyberstalking within the data sets. Furthermore, the reasoning as to why the data set were used and results are discussed in greater detail alongside: graphs, charts, and RStudio codes and findings. In addition, the limitations and recommendations for future studies are suggested and mentioned, as well as what are expected to achieve and potentially continue to do after this study is concluded. Finally, the next step is to explore the data collection and analysis methods, using algorithms and focusing on the metrics of cyberstalking.

Chapter 5: Use of the K Means Clustering Algorithm to Analyse Twitter data

5.1. Introduction

This chapter presents K means clustering algorithm on the random sample data sets that were used and mentioned in the previous chapter(s). The k means clustering algorithm is generally the most known and used clustering method. The reason as to why this method was used was for testing a few of the research questions and its aims. There are various extensions of k means to be proposed within literature. Although it is an unsupervised learning to clustering in pattern recognition and machine learning, k-means algorithm and its extensions are always influenced by initializations with a necessary number of clusters. In addition, within this chapter another main focus is on the results of k-means; as well as the liking between those results and cyberstalking indicative content on Twitter.

Therefore, in this chapter, tweets that were extracted, refined, analysed, and visualised for representation. Are refined even further with k means its results. A goal for this research is to visualise the cyberstalking tweets in a particular area, on Twitter and visualise the clustered emotional terms according to the liking or connection to cyberstalking indicative content on Twitter. As well as, how the random sample data set provided helps illustrate the correlation between unigrams and emotional terms towards cyberstalking suggestive content on Twitter. Lastly, many of the various packages and libraries are provided by R for extracting and processing the data and also for the visualisation of clustered data. As well as, within this chapter, the researcher is concentrating on the Twitter data from the random sample data set. Also, R language is used for acquisition, pre-processing, analysing and visualization of the

twitter data specified. Thus, taking the Twitter data that was extracted and analysed, then pre-processed and now will be clustered.

A fundamental problem that habitually arises in a great variety of fields such as pattern recognition, image processing, machine learning and statistics is the clustering problem (Jain, A.K., Murty, M.N. and Flynn, P.J., 1999). In its basic form the clustering problem is defined as the problem of finding homogeneous groups of data points in a given data set. Each of these groups is called a cluster and can be defined as a region in which the density of objects is locally higher than in other regions. The simplest form of clustering is partitional clustering which aims at partitioning a given data set into disjoint subsets (clusters) so that specific clustering criteria are optimized. The most widely used principle is the clustering error principle which for each point computes its squared distance from the corresponding cluster centre and then takes the sum of these distances for all points in the data set. A popular clustering method that minimizes the clustering error is the k means algorithm.

However, the k means algorithm is a local search procedure and it is well known that it suffers from the serious drawback that its performance heavily depends on the initial starting conditions (Pena, J.M., Lozano, J.A. and Larranaga, P., 1999). Nonetheless, to treat this problem several other techniques have been developed that are based on stochastic global optimization methods (e.g. simulated annealing, genetic algorithms). However, it must be noted that these techniques have not gained wide acceptance and in many practical applications the clustering method that is used is the k means algorithm.

Furthermore, technology and information are becoming more and more increasingly sophisticated. Various agencies or organisations manufacture and accumulate large amounts of data in their database. The most prevalent technique

used to obtain any database or big data is called data mining. Simhachalam & Ganesan (2016), suggests: data mining is defined as an analysis process to find valid and unexpected relationships between data sets and convert data into data structures so that they are easy to understand and useful for users. Data mining analysis techniques generally consist of prediction techniques, description techniques and inference techniques. Grouping is one of the description techniques of data mining analysis. For instance, in general there are two methods of grouping, the first method is hierarchy, and the second method is non-hierarchy.

As mentioned before, one of the most popular or known non-hierarchical clustering methods used is, the k means method. K means is also known as hard clustering which can group objects with clear boundaries, meaning that they can group objects into certain groups and not members of other groups (Sivarathri & Govardhan, 2014). The k means method is a partition-based method that attempts to partition data into two or more groups using the mean value as the centre of the cluster. Moreover, it is important to know with the k means method there is also the k medoids method which is a partition-based method that uses medoids as the centre of the cluster. Medoids is the most centralized cluster data object (Arora et al., 2016), so this method is more robust to outliers than the k means method (Tiwari & Singh, 2012).

In addition, brought forth by Lloyds, 1982, introduces what K means algorithm is and how it works. K Means (Lloyd, 1982) K Means is the widely used iterative clustering algorithm. As the algorithm requires the number of clusters K to be provided as input. It works as follows:

a) Initialise the centroids randomly.

b) Compute the distances of the data points from each of the centroids.

c) Assign the data point to the closest centroid.

d) Update the centroids.

e) Repeat the steps from (b) to (d) for the desired number of iterations.

A social structure of individuals related directly or indirectly on the basis of some common factor like similar likings or retweets, is a social network. In order to understand the behaviour and structure of a social network we need to study the network and this study is called social network analysis. There has been a rapid increment in the research and study of data mining community and social network analysis (Garg and Rani, 2017). There is a vast quantity of social networking sites available on Internet such as: LinkedIn, Facebook, Instagram, Twitter, Google and many more. However, with the many interactions over such sites produces such huge amount of data because billions of active users maintain their accounts. Hence, it is a tedious task to analyse the complex data. It is of great importance for academic and business to analyse such online social communities and predicting their behaviour.

Additionally, this chapter will be shown in detail how to construct an unsupervised learning algorithm: k means algorithm, from the data sets collected and mentioned previously. This is done so that the data sets are free of initializations without parameter selection and can also simultaneously find an optimal number of clusters within the tweets. Throughout this chapter, each section will go into further detail about: the twitter threads formerly collected, the unigrams, and lastly, the emotional terms that have been reoccurring and what clusters they were collected in and the association between the two and correspondence or relationship between cyberstalking.

5.1.1 Social Media: Twitter and K-Means

5.1.1a Twitter

The past decade has seen a rapid expansion of social networking platforms, along with the users that use them. Online social networking sites not only connect people, but these platforms also allow users to discuss their opinions related to any political, social, or everyday matter. In addition, while connecting people from different parts of the world, they can raise their voices in favour of or against any global, national, or even local issues that are being narrated online. Also, with the rapid development of social networking platforms comes the extension of digital technology, they go hand in hand. Furthermore, with the everlasting digital age, important professional means, policymakers, and even law enforcement can get useful insights into public opinion from all or any social networks.

One of the most popular online social networks is Twitter. Twitter is an American microblogging and social networking service on which users post and interact with messages known as "tweets". Registered users can post, like, and retweet tweets, but unregistered users can only read those that are publicly available. According to the 2019 statistics (taken from https://blog.hootsuite.com/twitter-statistics/), around 326 million people use Twitter every month. As well as 500 million tweets are sent every day which means 5787 tweets are posted every second. The platform allows users to post a short message or tweet in 34 different languages which include and are not limited to: Arabic, English, Bengali, Chinese, French, Spanish, Urdu, German, Russian, and countless others. People use Twitter frequently to express their opinions on government initiatives, societal problems, religious affairs and a widely catalogue of supplementary topics. However, the enormous amount of Twitter messages produced each day makes it unmanageable to manually process a tweet to establish the topic of public discussion. In addition, people use various

hashtags to express their own opinion on the same topic. Which also allows an individual to reach more users using a hashtag rather than a select few. This makes the task of grouping semantically coherent tweets more challenging.

Moreover, as stated above, Twitter is a social networking service which allows the user to send and read the short message of 140 characters called “tweets”. For instance, there are two types of users for twitter account holders. As stated above, one is registered users who can only read the tweets, and another are registered users who can read and post the tweets themselves. Twitter is a public platform for all the people of different age categories all over the world to connect with one another. Moreover, the data generated by Twitter is heterogeneous in terms of content because user can post a text, image, video, and audio in any format. Likewise, data is also big in size because hundreds of thousands of tweets per day is generated (A. Sechelea, T. Do Huu, E. Zimos, and N. Deligiannis, 2016). Moreover, which is relatively new for the platform Twitter in the late 2009, twitter added a new feature which allows each tweet to be geo-tagged which is associated with longitude and latitude of specific location. As seen, in this chapter the tweets that are being used were extracted previously, refined, analysed, and then visualised.

Likewise, social media is an Internet-based application built with Web 2.0 technology and allows the exchange of user-generated content (Kaplan & Haenlein, 2010). One of the most popular social media networks right now is in fact Twitter. Twitter is used daily to exchange ideas, gather information, and see the activities of users that are followed (Java et al., 2007). Ideas sent via Twitter by many individuals who use the social media site are called tweets. Tweets are stored in the Application Programming Interface (API) feature that can be accessed by users. The desired information can be found based on the keywords entered so that the tweets obtained

are in accordance with the topics discussed. For instance, let's presume the keyword used in this case is "cybserstalking", then tweets that have the word "cyberstalking" will be picked up by the system. When withdrawing Twitter data, anyone needs to get permission from Twitter to get the API access code. In order to get the access code, it has to be coming from a registered Twitter account requesting the access. Traditionally, there are four access codes, namely consumer key, consumer secret, access token, and access secret. Data withdrawal can be done if you already get the access code by integrating Twitter API and RStudio. During data withdrawal the Internet connection must always be activated. Lastly, as mentioned Twitter is an online news and social networking site where people communicate in short messages called tweets. There is another description of Twitter and tweeting might be microblogging which is described and detailed below. Interestingly enough some people use Twitter to discover interesting people and companies online, opting to follow their tweets thus starting the process of microblogging.

5.1.1b Microblogging

Furthermore, Twitter is one the most used microblogging site within social media currently now. However, the difference between a blog which is simply a web page that contains informational posts by one or multiple users, often related to a specific topic. Likewise, as comparison to microblogging, on the other hand, refers to short messages or posts shared with an audience online through Microblogging platforms, such as Twitter, Instagram, and Tumblr for instance. As for the reasoning to why microblogging is becoming and has become a huge success, microblogs allow users to exchange small elements of content such as short sentences, individual images, or video links, which may be the major reason for their popularity. However, another form of microblogs exists like commercial microblogs which endure to promote

websites, services, and products and to promote collaboration within an organisation. The Twitter microblogging system burst into public view only a few years ago and has now become a worldwide phenomenon. It is like traditional blogs in its focus on recent posts, but differs in that its posts, called “tweets”.

Moreover, Twitter takes advantage of the idea of blog feeds by allowing you to subscribe to, which is, the “follow” indication on Twitter which allows an individual to follow any other Twitter user. A user's personalized feed shows the most recent tweets of all individuals he or she is following, creating a live stream of bite-sized information pieces. Many several competing services exist, some of which overlay the service on top of other services. For example, other important social media networks such as Facebook and LinkedIn have status messages that serve as microblogs that are broadcast to friends. More recently, Google Buzz and the open source identi.ca provide similar services to any of their account holders.

Viewing how and what microblogging sites offer to the public is vast and endless which help create several interesting social network structures. The most obvious network is the one created by the “follows” and “is followed by” relationships which the main function of Twitter. However, unlike Facebook, these “follow” relationships are potentially directed: you can follow people who don't follow you and vice versa. This contrasts with the undirected ties present in the other examples stated beforehand, Facebook and LinkedIn. Other networks are created that connect all the users together based on the number of times they reply to others' microblog posts or repost messages they come across.

Additionally, as is important to know, that all social media networks provide a vast and large amount of data. Automation is required for extracting the knowledge from large volume of data. It is a challenge for both the developer and algorithm to

compute the data quickly. Recently, microblogging has become a popular trend which is responsible for a large amount of information dissemination. Microblogging websites are services which allow the user to post their ideas, opinions in defined number of words. It also allows the user to exchange content, images, video links and others. However, microblogging sites have become popular due to rapid growth of Twitter. As mentioned before, other microblogging sites are also available with the same functionality such as: Tumblr, Pinterest, Flattr, and Plurk, and many others. Moreover, Tumblr provides the same functionality as Twitter; however, it focuses on the design and style. It is best for its simplicity in content management and posting. Each Tumblr blog is known as tumblelog. The important and powerful feature of social networking and microblogging platform is that user can post a message only to a selected friend or group of friends and not necessarily for all friends.

5.1.1c K-means clustering

The K-means clustering algorithm is used to find groups which have not been explicitly labelled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets. In addition, clustering algorithms are very beneficial and useful tools for data mining. Compression, probability density estimation, and many other important tasks. However, most clustering algorithms require the user to specify the number of clusters (called k), and it is not always clear what the best value for k. Choosing k is often an unplanned decision based on prior knowledge, assumptions, and practical experiences. Choosing k is made more difficult when the data has many dimensions, even when clusters are well-separated. Centred-based clustering algorithm in particular k means and others, usually assume that each cluster adheres to a unimodal distribution. Whit these methods, only one centre should be used to model each subset

of data that follows unimodal distribution. If multiple centres are used to describe data drawn from one mode, the centres are needlessly complex description of the data, and in fact the multiple centres capture the truth about the subset less well than one centre (Hamerly, Elkan, 2004).

Likewise, the performance of a clustering algorithm may be affected by the chosen value of k, as stated above. Therefore, instead of using a single predefined k, a set of values might be adopted. It is important for the number of values considered to be reasonably large, to reflect the specific characteristics of the data sets. At the same time, the selected values must be significantly smaller than the number of objects in the data sets, which is the main motivation for performing data clustering. Many reports or studies done on or with K means clustering and its applications usually do not contain any explanation or justification for selecting values for K.

The clustering algorithm that is the focus of this study again is K means algorithm. The K means algorithm accepts two parameters as input:

- The data;
- A K value, which is the number of groups that the researcher or an individual wants to create.

Conceptually, the K means behaves as follows:

I. It chooses K centroids randomly;

II. Matches each point in the data (in some cases, each mammal) with the closest centroid in an n-dimensional space where n is the number of features used in the clustering (for example, 5 features: water, protein, fat, lactose, ash). After this step, each point belongs to a group.

III. Now, it recalculates the centroids as being the mean point (vector) of all other points in the group.

IV. It keeps repeating the steps 2 and 3 until either when the groups are stabilised, that is, when no points are reallocated to another centroid or when it reaches the maximum number of iterations (the stats library uses 10 as default). A pronounced package within R that introduces a user to k means is: tl;dr

The below steps serve as a starter to the k-means clustering method.

1. Data Preparation: Preparing your data for cluster analysis
2. Clustering Distance Measures: Understanding how to measure differences in observations.
3. K Means Clustering: Calculations and methods for creating K subgroups of the data.
4. Determining Optimal Clusters: Identifying the right number of clusters to group your data.

The processing of the data is very important as well for instance, to perform a cluster analysis in R, generally, the data should be prepared as follows:

1. Rows are observations (individuals), and columns are variables
2. Any missing value in the data must be removed or estimated.
3. The data must be standardized (i.e., scaled) to make variables comparable. Recall that, standardization consists of transforming the variables such that they have mean zero and standard deviation one (UC Business Analytics R Programming Guide, Web, 2018).

Furthermore, therefore choosing a good k is essential, the bigger is the k you choose, the lower will be the variance within the groups will be in the clustering. If k is equal to the number of observations, then each point will be a group and the variance will be 0. It's interesting to find a balance between the number of groups and their variance. A variance of a group means how different the members of the group are. The bigger is the variance, the bigger is the dissimilarity in a group. As well as, sorting your data properly and having it ready for the k-means clustering algorithm.

5.2. K-means Clustering in R Programming

For instance, the use for a clustering algorithm is because, cluster analysis is a statistical method which aims to classify several objects into some groups (clusters) according to resemblances between them. While it has been widely used in many purposes such as DNA microarray analysis, the uncertainty of results caused by sampling error of data has not generally been evaluated in practice. Pvclust is an enactment of bootstrap analysis on a statistical software R to assess the uncertainty in hierarchical cluster analysis.

The importance of uncertainty assessment has been sound and recognized in phylogenetic analysis. It is a special form of hierarchical clustering for inferring the history of evolution as a dendrogram. Thousands of bootstrap samples are generated by randomly sampling elements of the data, and bootstrap replicates of the dendrogram are obtained by repeatedly applying the cluster analysis to them (Efron, 1979; Felsenstein, 1985). The bootstrap probability (BP) value of a cluster is the frequency that it appears in the bootstrap replicates. The multiscale bootstrap resampling was developed recently (Efron et al., 1996; Shimodaira, 2002, 2004) for

calculating approximately unbiased (AU) probability values (p-values) as implemented in software consel (Shimodaira and Hasegawa, 2001).

Also, there has been an assortment of algorithms has been industrialised in the past decades to improve performance associated to the primitive algorithmic setup, in particular Anderberg (1973, page 135), Rohlf (1973), Sibson (1973), Day and Edelsbrunner (1984, Table 5), Murtagh (1984), Eppstein (2000), Cardinal and Eppstein (2004). Also, hierarchical clustering methods have been implemented in standard scientific software such as R (R Core Team 2011), MATLAB (The MathWorks, Inc. 2011), Mathematica (Wolfram Research, Inc. 2010) and the SciPy library for the programming language Python (Jones, Oliphant, Peterson et al. 2001; van Rossum et al. 2011). Specifically, there are the following functions:

- hclust in R's stats package (R Core Team 2011),
- flashClust in R's flashClust package (Langfelder 2011),
- agnes in R's cluster package (Ma¨chler, Rousseeuw, Struyf, Hubert, and Hornik 2011),
- linkage in MATLAB's statistics toolbox (The MathWorks, Inc. 2011),
- Agglomerate and DirectAgglomerate in Mathematica (Wolfram Research, Inc. 2010),
- linkage in the Python module scipy.cluster.hierarchy (Eads 2008).

Clustering is a board set of techniques for finding subgroups of observations within a data set. Within the cluster observation, ideally the observations would be favourable if they are in the same group to be similar and observations in different groups to be dissimilar; because there is not a response variable, this is an unsupervised method, which implies that it seeks to find relationships between the *n*

observations without being trained by a response variable. Clustering is used because, it allows to identify which observation are alike, and potentially categorise them therein. K means clustering is the simplest and most used clustering method for splitting a dataset into a set of k groups.

As mentioned before k means clustering is an unsupervised learning approach to machine learning. Clustering is a useful tool in data science, it is a method for finding cluster structure in a data set that is characterized by the greatest similarity within the same cluster and the greatest dissimilarity between different clusters. Hierarchical clustering was the earliest clustering method used by biologists and social scientists, whereas cluster analysis became a branch of statistical multivariate analysis (Sinaga and Yang, 2020).

K means clustering is extensively used in various fields such as text mining, machine learning, image analysis, image processing, web cluster engines, bioinformatics, weather report, and so on (Bijuraj, 2013). Hence, why k means is being used for this study. It has been shown that k means is used in various fields and two of those fields are the main methods of this research: text mining and machine learning. Making k means the seamless algorithm to carry out this study along with the methods that the researcher has selected for this study. Moreover, there are diverse methods of the clustering for instance, model-based method, density-based method, hierarchical method, grid-based method, partitioned method.

Nonetheless, k means is used for the unlabelled data i.e., data are not labelled into any group of clusters. The objective of this algorithm is to find clusters in the data with the already given number of clusters. The number of clusters and the dataset are the inputs of the algorithm. The dataset is the collection of the data for each data point. The number of clusters can either be randomly selected or randomly generated from

the dataset. The algorithm selected works as: firstly, initialise the number of clusters and the set the centroid of the clusters. Each data point is assigned to a cluster based on the smallest distance between the centroid and the data point. The centroids are updated or recomputed by taking the average (mean) of the data point assigned to the cluster. The process continues until stopping criterion is met. The stopping criterion is any one of them which are data points is not changing in the clusters, the sum of the distance is minimised, or the number of iterations has reached the maximum.

The main goal of the current research is to visualise the linking of the results from the k means cluster algorithm along with the unigrams; to showcase the correlation or relationship towards cyberstalking indicative content within tweets on the social media platform Twitter. As well as it envisions the clustered tweets or terms according to the relationship of cyberstalking. Moreover, the various packages and libraries are provided by R for extracting and processing the data and for the visualisation of clustered data will be mentioned and shown as previous chapters.

5.2.1 Clustering Programming

K Means Clustering in R Programming is an Unsupervised Non-linear algorithm that cluster data based on similarity or similar groups. It seeks to partition the observations into a pre-specified number of clusters. Segmentation of data takes place to assign each training example to a segment called a cluster. The advantage of using k means in R is profound. R gives any researcher or individual a great base to conduct their analysis on his or her data. How to use k means clustering in RStudio, as mentioned in pervious chapters the dataset needs to be imported into RStudio. Once the dataset is in RStudio and cleaned up, if needed you can begin to use k means clustering. Again, k means will cluster that data into groups, that data itself will not be pre-grouped. K means will determine how many groups or clusters are within the

dataset. This process was done to each random sample data set (5 csv files) within RStudio.

However, this process was the final set within R Programming. As previously stated, R was used for the data mining and analysis of the collected data for this study. Within past chapters and mentioned in their selected sections R was used for many purposes. K means clustering was done at the final stage after importing data sets, cleaning the data sets, removal of information or characters not needed, turned into a matrix, and then graphed, and lastly K means was conducted. Likewise, just as before, K means was conducted on each of the five data sets separately in RStudio. Therefore, the researcher has five different K means clustering results for each data set. These results consist of dendrogram charts and results within the console broken down to the mean, median, and mode of the cluster for each data set.

Nevertheless, like all machine learning and data mining programmes there are some limitations that each programme or algorithm might have. For instance, it is important to know and understand there is a limitation to using k means. The most important limitations of Simple k-means are: The user must specify k (the number of clusters) in the beginning. k-means can only handle numerical data. k-means assumes that we deal with spherical clusters and that each cluster has roughly equal numbers of observations. Therefore, as the researcher as stated and mentioned previously, that specifying the number of clusters at the start, as well as the k means can only handle numerical data. Hence, why this dataset was used for the purpose of this algorithm.

For example, below is the entire console for the k means clustering algorithm that was done within RStudio for the random sample data set: dataset 1 csv file. At

the top of the results from the console, which shows the stringrs that were used to conduct k means in R.

- hc <- hclust(distance, method = "ward.D")
- plot(hc)
- #nonhierarical k-means
- m1 <- t(m)
- set.seed(222)
- k <- 12
- kc <-kmeans(m1, k)
- kc

The above strings were used in R towards each data set. The set.seed() function sets the starting number used to generate a sequence of random numbers. This is done to ensure that you or the individual using this stringer is to get the same result if you start with that same seed each time you run the same process. Set the seed of R ‘s random number generator, which is useful for creating simulations or random objects that can be reproduced. The random numbers are the same, and they would continue to be the same no matter how far out in the sequence the researcher went. A great tip for future endeavours, use the set.seed() function when running simulations to ensure all results, figures, are reproducible. The k, is set to 12, because having twelve clusters is more than substantial for this research. The m1 is the matrix that was made and used in the previous chapter along with the k that was just set to twelve. Lastly, the entire algorithm is run, hence, kc.

Once the algorithm is run within R, and the raw results are brought into the console, the researcher can look through the raw results and optimise the important aspects. For instance, as mentioned throughout this thesis, each data set was

analysed in R, therefore, after the clustering algorithm was done for each data set separately, each raw result was analysed and gone through in detail. Each raw result from the data sets were put into a separate word document and examined for which clusters are more prevalent in comparison to others. The researcher noticed similarities within the findings and the results to previous chapters. For example, three key words have been flagged or highlighted during the clustering algorithm process. However, they have also been seen throughout the entire data collection process: in both the preliminary data set and the random sample data set; and within the R programming analysis process on the random sample data set. In addition, these three key terms were also brought forward within Chapter 4, during the advanced search in R.

5.3. Reoccurring Themes: Preliminary and Random Sample Data Set(s)

The Preliminary data set and the random sample data set were both used throughout this thesis and highlighted in their perspective chapters. In addition, again the random sample data set was used in for the clustering process done within this chapter. As stated before, while the random sample data set was used within R Programming many (other) functions were done and are highlighted in other chapters of this thesis. For instance, during the data collection and analysis process reoccurring themes and key words were showcased throughout. While the researcher noticed from the start while he or she obtained the preliminary data set that there is in fact reoccurring themes and key words reoccurring, which are being referred to as emotional linking's (in this chapter), that are focusing on cyberstalking on Twitter.

As the researcher looked over and examined the data set that was collected for the preliminary portion of this thesis. He or she spotted emotional terms or key words that were reoccurring regarding the topic at hand all lining towards cyberstalking

suggestiveness on Twitter. With that being stated, there are in fact (emotional) reoccurring themes and key words that are identified along with the unigrams that were used for the motivation this research. For example, while obtaining the countless tweets and feeds in the first stages of the preliminary data collection, the researcher noticed emotional linking or key words appearing in the word frequencies that correlate with the unigrams that he or she was looking into. As stated in chapter four of this thesis, the researcher supposed: the word frequency that was conducted on the random sample data set results were more of emotional terms within the usage of everyday terms in tweets. In fact, the researcher stated in Chapter 4: “that there is a similarity to emotional terms that can be used within or as an exception to cyberstalking”. Therefore, the data sets along with the unigrams provided within this research have numerous links, reoccurring themes, key words, and emotional ties towards cyberstalking allusive content on Twitter.

Furthermore, as the researcher stated early on in this chapter. That from the beginning of the data collection process, he or she noticed emotional terms or key words, side by side amongst the unigrams within tweets. However, lining of the emotional terms had to be verified or proved within the random sample data set. Nonetheless, before that is done, here are fifteen tweets with each unigram and the three emotional terms the research focuses on for this chapter. Below within the tweets are the unigrams highlighted in yellow and the three emotional terms: bad which is highlighted in green, sad which is highlighted in blue, and hate which is highlighted in pink. The blacked-out parts of the tweets are usernames, people’s names, or personal information that is being kept anonymous for the purpose of this study. These fifteen tweets are tweets that were obtained during the preliminary data collection. All of the fifteen tweets that are shown beneath are from the 5,000 tweets used in chapter three.

Tweets being used are below:

- Stalking is also abusive. If your partner stalks you, it's wrong. There are laws that govern these actions and they are trying to expand them more to address the nuances of modern culture all the time (I.e. technology). Its sad and frighting but you are not alone.

- Why does every guy I talk to end up being an annoying creepy stalker I fucking hate it

- [redacted] Sounds like a creep. Too many of those around. They'll stalk your every post on social media. Its sad and disgusting.

- I won't stop speaking out and I do not deserve to live my life in fear. #cyberstalking is a hateful crime it needs to be stopped

- Once a man whistled at me and I told him I'm not a dog. He then proceeded to call me a fucking bitch and followed me down the length of abbey street until I ran into bus aras. Whistling is the tip of the iceberg. What lies beneath is following, verbal abuse, attacks, stalking its awful and sad that this happens

- focus on the bigger picture here. There are bigger issues than disrespect or hate. Rape, child abuse, harassment and stalking, acid attacks etc. these are all violent bad horrific crimes

- Maybe someone I know because immediately after adding me I got a message… you know the creepy kind like I said before maybe someone I know that can be trying to stalk me. You know women and stalking it's an actual fear and I hate it so much so after seeing it I went and blocked

- This creepy man has been stalking me for the longest on Twitter. I'm just now saying something because I'm starting to really hate it and im getting scared now.

- toxic ex best friend (relationship) that used to stalk me on insta tried to follow both of my accounts again so i removed her as a follower and went private also she's using an old account of hers and a different name so i won't know it's her but i remember that acc, you aren't clever [redacted] you are sad xx

- [redacted] helloo, i guess i have a stalker. pls ur obsession is scaring me i hate it

- The real #trauma from stalking has nothing to do with the bad actions the stalker takes, but the way they shift the victim's perspective forever scared and terrified.

- Another guy I blocked for bad toxic trolling turned out to be a complete creep and misogynist. Just another day on twitter oh technology.

- Having a stalker is bad its so annoying and unwanted like just leave me alone already

- I don't think y'all understand how bad it is for women on social media or in life. How many times do we have to hear about women being murdered for saying no. Women being stalked. Women going to police only for nothing to be done about their stalker or rapist or violent ex bf.

- I'm going through a bad/scary situation with my ex. I want y'all to know id NEVER kill myself. If I end up dead it'll be at the hands of my ex[redacted]. He's been stalking me, slashing my tires, harassing me and the police will do nothing until he touches me. I'm scared.

With having these tweets above help illustrate that the emotional terms are in fact impactive towards this research. For instance, these emotional terms were seen while preliminary data set was collected. Therefore, this notification of these terms early in the data collection process. supports that the emotional terms are amongst tweets associating cyberstalking indicative content on Twitter. Shown, in the figure below, which showcases the top three emotional terms that have been reoccurring during each data set analysis. The figure below, helps correlate the linking of the findings along with the tweets above. As you can see from the figure, Figure. 31, "sad", "bad", and "hate" are the three emotional terms that are highlighted. These three terms are a great focus for the researcher and this research. Mainly, because these three words are used in everyday language meaning that the results would vast and have standing in correlation to cyberstalking investigative content on Twitter.

Word Freq	Count	Rank
sad	1470	23

will	1453	24
night	1373	25
feel	1369	26
think	1365	27
need	1344	28
bad	1338	29
lol	1295	30
well	1286	31
sleep	1256	32
wish	1255	33
can	1240	34
didn't	1235	35
sorry	1234	36
morning	1202	37
much	1181	38
see	1164	39
hate	1048	41

Figure 32. Random Sample Data Set 3: Partial Word Frequency

The linking and the correlation between both data sets: preliminary data set and random sample data set are imperative towards the results within this study. Likewise, the researcher noticed similarities while obtaining the results for both data sets. However, the researcher needed to prove that these emotional terms are actually linked towards cyberstalking. Hence, why having the random sample data set is important for this research. Brining the random sample data set into this study will influence the connecting or relationship to the unigrams and emotional terms in correlation towards cyberstalking on Twitter. As mentioned before and throughout this thesis, the random sample data set is just that random, there is no distinctive correlation with regards to cyberstalking. Thus, using the preliminary data set findings: the unigrams, along with the random sample data set, each is vital towards this research. Making these results and / or findings within this chapter even more practical for this study. Proving that there is a sold connection between the data sets: preliminary data set and the random sample data set; alongside the findings / results,

which mainstream the suggestive or indicative content towards cyberstalking on Twitter.

5.3.1 Results from the Clustering Algorithm

The results for the random sample data set along with the use of clustering can be broken down into two different parts. For example, before the final results were acquired the researcher ran the clustering algorithm within R and was given the raw results or the output for each data set. Each raw result was broken down into twelve different clusters with the most common used terms in each cluster. For example, this is the raw results for the random sample data set: data set 3.csv file: K-means clustering with 12 clusters of sizes 1688, 3155, 1229, 2176, 1604, 2092, 2381, 3259, 27883, 2073, 1534, 926. Within these clusters the emotional terms: “bad”, “hate”, and “sad” all are within these clusters. “Bad” is in the first cluster, followed by hate which is in the fifth cluster, and lastly sad which is in the ninth cluster. Now, this was done to each data set and each data set had different results. Furthermore, data set 1 all three emotional terms were in one of the twelve clusters: “bad” again was in the first cluster, “hate” was in the fourth, and lastly “sad” was in the sixth. Data set 2 had same results all three terms were in one of the twelve clusters: “bad” again was in the first cluster, “hate” was in the eighth, and lastly, “sad” was in the fifth. However, for data sets 4 and 5 only two of the terms were in one of the twelve clusters. The two terms: “sad” and “bad” were seen in one of the twelve clusters; for data set 4 “bad” was in the first and sad was in the seventh and for data set 5 “bad” again was in the first and “sad” was in the sixth. It is very interesting to see that the term: “bad” was in the first cluster for each data set.

This correlates that the term itself is used more or brought up more in comparison to the other terms. Furthermore, the term “hate” was mixed throughout the clusters and again only seen in three out of the five data set results. While in comparison, the term “sad” was seen in all five data set results just like the term “bad”. However, the term “sad” was spread out within different clusters like the second term “hate”. Now, does that mean the two terms “hate” and “sad” are not as commonly used within the random sample data set in comparison to the term “bad”. Moreover, making “bad” and “sad” the two terms that were in all five data sets and “hate” only being three out of the five. Therefore, making “sad” and “bad” displayed 100 percent in comparison to “hate” with 60 percent. Lastly, remembering that the random sample data set does not have any indefinite original correlation to cyberstalking. Therefore, making the results and findings that more imperative to the cyberstalking indicative content that is on Twitter.

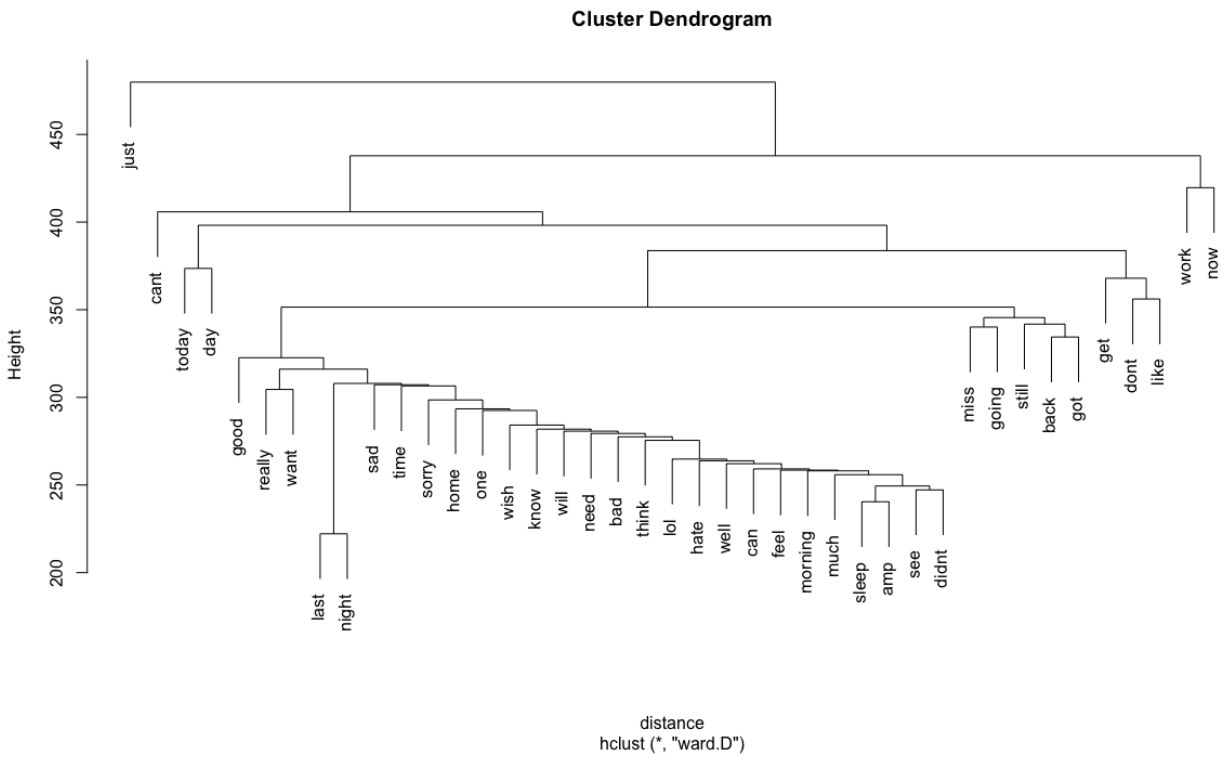


Figure 33. Cluster Dendrogram made in R Programming: details the Random Sample Dataset, dataset 3

As illustrated in Figure 32, “bad”, “sad”, and “hate” are all shown within the dendrogram. “Sad’ has the highest rank in comparison to the other two terms and “bad” and “hate” are closer together on the contrast. These terms are constantly showing persistence throughout each data collection and the analysation process. Which will be mentioned in even more detail with the visual from the tweets taken from the advanced searched results from chapter four. Also, there are other terms that one might think can be useful, however, for the purpose and time limitation of this study these three are the focus.

Below seen in tables 2-6, are the cluster results for the data sets for the random sample data set (all five). Each data set is broken up into its own table with each emotional term and the numeric value of the clusters they are in. Followed by the sum and average of each then tallied up and totalled together to get the overall value of each term within the random sample data set. This process was done after the raw data result from each data set was obtained. The researcher put each data set into excel and manually entered a made table by the researcher; then once finished each emotional term was tallied together with its cluster for each data set.

Figure 34:Table(s) 2-6: Cluster Results, using K-Means Clustering Algorithm

DATA SET 2		
Hate	Sad	Bad
0.01943095	0.40700902	0.01283831
0.0290631	0.01912046	0.04359465
0.01572327	0.02410901	0.03878407
0.02184525	0	0.02423403
0.03391473	0.0125969	0.02713178
0.02402521	0.0401733	0.0303269
0.02529602	0.02637244	0.02206674
0.01729631	0.03777879	0.03504779
0.02130786	0.01763409	0.0286554
0.02132867	0.04125874	0.02447552
0.01740812	0.03481625	0.03191489
0.02075154	0.01346046	0.0330903
0.26739103	0.67432946	0.35216038

DATA SET 3		
Hate	Bad	Sad
0.008293839	0.04028436	0.01777251
0.000633914	0.03074485	0.03042789
0.013018714	0.02847844	0.03498779
0.002757353	0.01930147	0.02757353
0.006234414	0.02119701	0.01745636
0.005258126	0.0540153	0.05736138
0.005039899	0.0247795	0.02183956
0.001534213	0.02516109	0.02117214
0	0.02492558	0.03066385
0.004341534	0.02074288	0.02556681
0.000651891	0.03389831	1.03887689
1.03887689	0.01943844	0.010799136
0.047763897	0.34296723	1.334497846

DATA SET 4	
Sad	Bad
0.035468502	0.02593965
0.028345144	0.02626094
0.008782936	0.03513174
0.034001214	0.02550091
0.043912888	0.01535166
0.032078964	0.50215916
0.027950311	0.03354037
0.066528067	0.02945253
0.040899796	0.01840491
0.048007838	0.02384063
0.013583138	0.01639344
0.036480532	0
0.41603933	0.75197594

DATA SET 5	
Bad	Sad
0.02593965	1.037098791
0.02626094	0.028345144
0.03513174	0.008782936
0.02550091	0.034001214
0.01535166	0.043912888
0.50215916	0.032078964
0.03354037	0.02542726
0.01834862	0.066528067
0.01840491	0.040899796
0.02384063	0.048007838
0.01639344	0.013583138
1.035429584	0.036480532

1.776301614	1.415146568
1.776301614	1.415146568

DATA SET 1		
Sad	Bad	Hate
0.01777251	0.0288124	0.008293839
0.03042789	0.0243967	0.000633914
0.03498779	0.0225829	0.013018714
0.02757353	0.0199653	0.002757353
0.01745636	0.028821	0.006234414
0.05736138	0.0215311	0.005258126
0.02183956	0.0160681	0.005039899
0.02117214	0.0286576	0.001534213
0.03066385	0.0252646	0
0.02556681	0.0265722	0.004341534
0.0273794	0.0338164	0.000651891
0.0237581	0.0181159	1.03887689
0.33595932	0.2946042	1.086640787

In the above tables are the three emotional terms that the researcher thought are the most beneficial for this study: “bad, “sad”, and “hate”. It is likely that, the unigrams are only part of the solution. Therefore, there needed to be another term or key word that can help the bring forth any cyberstalking indicative behaviour or content on Twitter. The researcher conducted the clustering algorithm k means for each data set and saw parallels within the raw results. Once, the raw results were then distributed into separate parts and the researcher determined all the major key themes; the researcher then broke the results into tables for each data

set.

The final tallied cluster results for data set 1:

Sad	Bad	Hate
0.33595932	0.2946042	1.086640787

The final tallied cluster results for data set 2:

Hate	Sad	Bad
0.26739103	0.67432946	0.35216038

The final tallied cluster results for data set 3:

Hate	Sad	Bad
0.047763897	0.34296723	1.334497846

The final tallied cluster results for data set 4:

Sad	Bad
0.41603933	0.75197594

The final tallied cluster results for data set 5:

Bad	Sad
1.776301614	1.415146568

Looking at these results: data set 1 the strongest term is: “hate”. In the second data set, data set 2: the strongest term is: “sad” and for data set 3: the strongest term is: “bad”. Finally, for the last two data sets, data set 4: the strongest term is: “bad” and for data set 5: the strongest term is: “bad”. From these results the emotional term that is constant is the term: “bad”. Now, as the researcher stated before the emotional term “bad” was seen in the first cluster of out twelve within the raw results regarding all five data sets. Then the average or sum of the terms within their clusters are highlighted in yellow at the bottom of each table. Thus after, the amicable results were concluded the sum of each emotional term was then presumed.

Cluster Results

Sad	Hate	Bad

0.33595932	0.26739103	0.35216038
0.65432946	0.047763897	0.75197594
0.41603933	1.086640787	1.776301614
1.415146568	1.401795714	0.29460419
1.334497846	0.467265238	0.34296723
4.155972524		3.175042124
0.831194505		0.703601871

Figure 35. Sum/Average Cluster Results: from all five datasets taken from the Random Sample Dataset

In Figure 34, the three emotional terms are shown with the sum of the total of each cluster within each data set, which is highlighted in yellow. As well as the average of the total the sums of each cluster from each data set, which is highlighted in orange. As previously stated, the emotional term “bad” was constantly in the first cluster within all five data sets. Therefore, making these results surprising to be the second strongest sum or value of the three terms with: 3.175042124 followed by “hate” with: 1.401795714 and lastly the strongest sum or value emotional term: “sad” with: 4.155972524. Thus, having the terms values from: “sad”, “bad”, “hate”. The two figures below, Figure 34 and Figure 35, are a better visual of the above figure, Figure 30. Both figures showcase the trajectory of each term within each data set and well as its cluster.

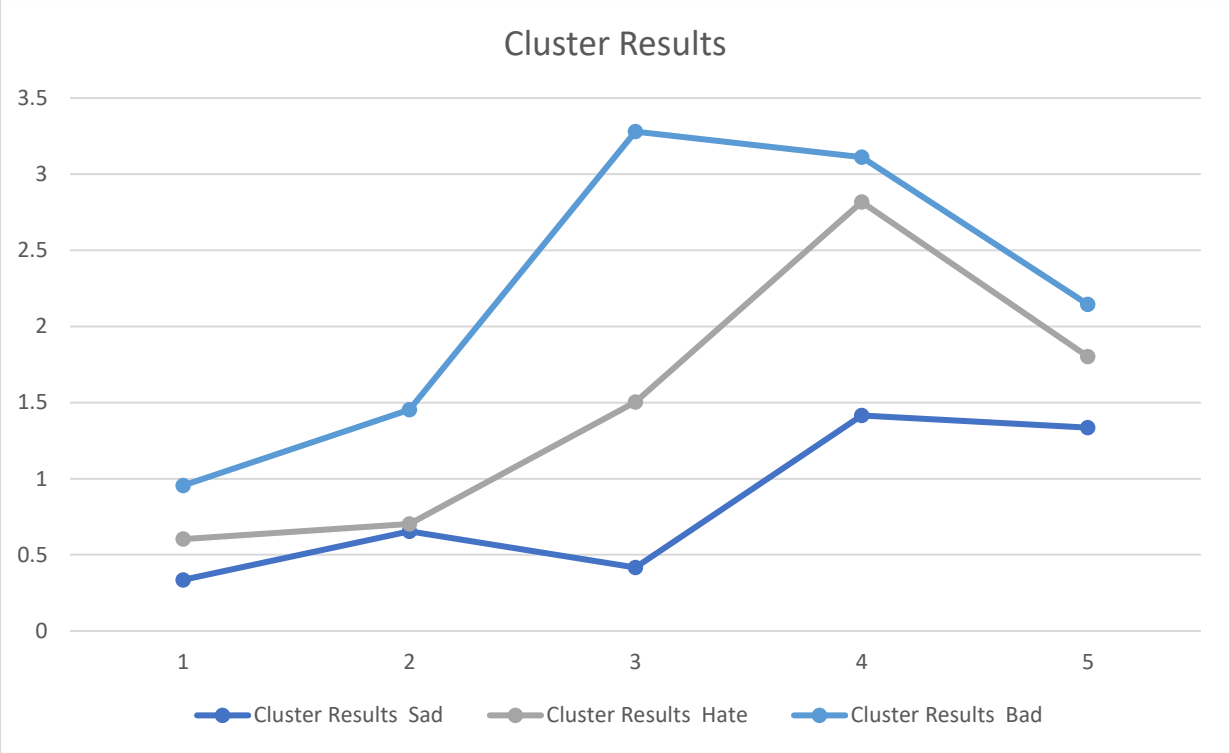


Figure 36. Line Graph from the K-Means cluster results used on all five Random Sample datasets

As seen in the above figure, Figure, 35 this chart shows the visualisation of each emotional term(s) side by side. Seen in the overhead figure the term “bad” has a high trajectory and then descends and levels out within the numeric values within the clusters it is represented. As for “sad” the trajectory this term seems to have high peaks and low valleys making it a constant term throughout the data set. Finally, the term “hate” starts off with a descend and the rises to a high peak, but then tappers off a bit and descends again. This visualisation of the trajectory and descends of each term are apparel for this research. It brings forth more concrete findings towards cyberstalking indicative content on Twitter.

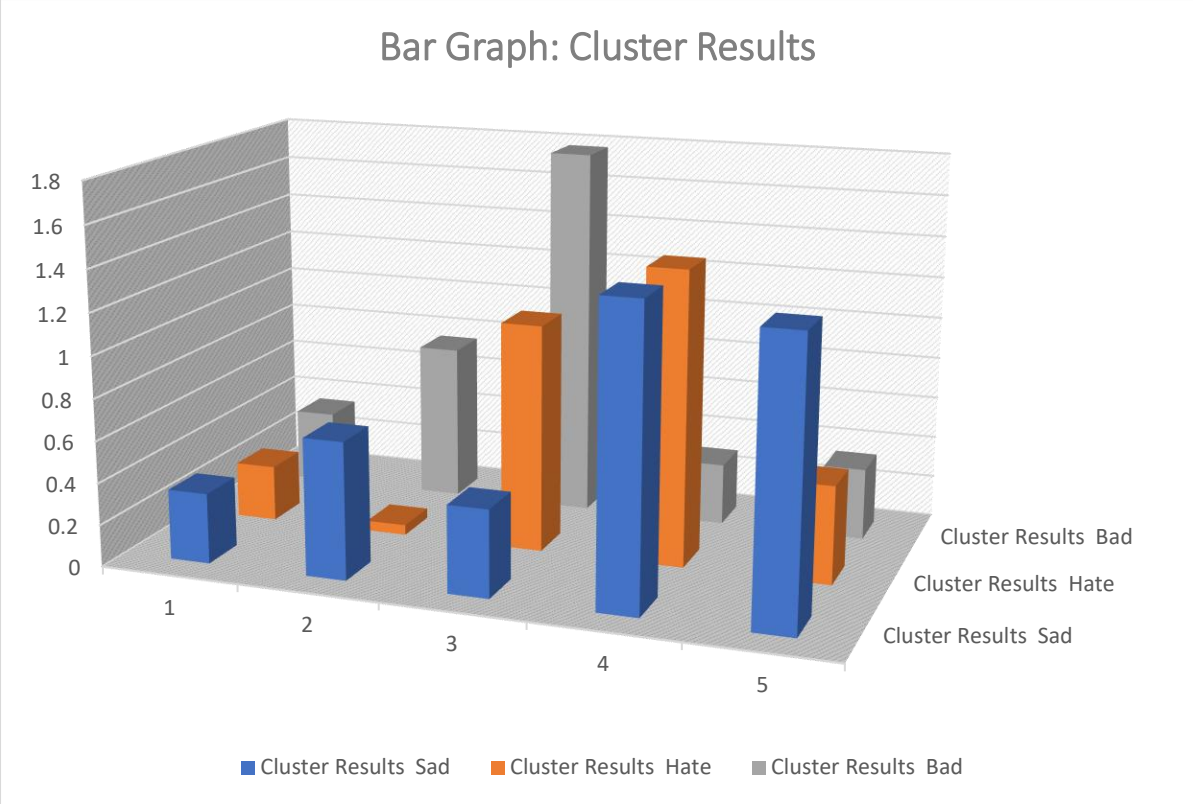


Figure 37. Bar Graph: of the K-Means cluster results taken from all five Random Sample Datasets

In the above figure, Figure 36, this bar graph gives another perspective into the clustering algorithm results. Not like seen in Figure 35, the results are parallel with one another and are shown the correlation within each data set and cluster. The bar graph which holds massive impact on the insight of these results are just as astounding as the line graph above, breaking down the results in a clearer image.

As you can see from the above figures, these three emotional terms are reoccurring through the random sample data set. The two terms “sad” and “bad” are present in all the five data sets and show high patterns of reoccurring trajectory. However, “hate” is not nearly as present within the random sample data set in comparison to the other two terms, but it’s a term to focus on for cyberstalking indicative content on Twitter. Another important finding is how these emotional terms are used in tweets on the social media platform Twitter that help showcase cyberstalking indicative content.

5.3.2. Clustering Results within Tweets

The prominence or understanding of how these three emotional terms and the fifteen unigrams showcase that there is cyberstalking indicative content on Twitter. The unigrams and these emotional terms go hand and hand with each other. Both findings are imperative towards the research at hand, however, they both are very impactful with regards to cyberstalking suggestive content either together or apart but more so together. Now this research needs to be imperative to the fact that cyberstalking indicative content is in fact on Twitter. Also, how that content can be detected and how it can possibly even be prevented. However, from the beginning while the collection of the preliminary data set was underway; the researcher kept in his or her mind that there had to be more than the unigrams that he or she suggests that are impetrative within this research. Yes, the unigrams are again a great starting point to prove that cyberstalking indicative content is in fact on Twitter. Conversely, not only do the unigrams that the researcher has advised and brought forward prove that cyberstalking suggestive content exists on Twitter, but that there is also an emotional link or an everyday use of terms with regards to the context of cyberstalking as well.

For example, while obtaining the preliminary data set the tweets that were collected had emotional tendency within the tweets itself. Which was shown above, fifteen tweets that highlight the unigrams and the three emotional terms. These reoccurrences from the preliminary data set were within each Twitter thread that was collected for the preliminary data set mentioned in chapter three. Below are ten tweets that show the unigrams as well as the emotional term which assist these results even further. Again, like above, the terms and unigrams are highlighted in each tweet. As you can see in these ten tweets: the unigrams are highlighted for the visual aid: the

yellow highlights are unigrams and the emotional terms are highlighted in different colours for clarity: hate is pink, sad is blue, and bad is green, as seen before. Also, for importance the blacked-out portions are someone's username or a name itself, again that is done for keeping this research anonymous.

- 12:20. Online at school. My classmate is a creepy fucking stalker, I hate him
- I DONT WANT YOU TO TALK ME TO ME YOU CREEP ME OUT STALKER MAKE ME SAD
- Fuck. Twitter. I'm actually fearing my life right now fuck cyberstalking ugh so bad
- [redacted] What do you mean Twitter isn't real life? Anyway it's sad she's stalking me, she's a bit obsessive like that.
- I dont think anyone understand the pain this is causing me I hate technology right now. fuck u twitter!
- Being hate, hate, hated on Twitter! Its sad and scary cyberbullying/ cyberstalking really fuck social media.
- [redacted] hope you aren't referring to me…id hate that say hello to your new stalker
- wondering how these freaks always seem to find me? so creepy...I mean seriously is their something that bad wrong with me?
- Twitter is pretty much legal stalking....And I like it...even though I do feel bad.
- Is getting abuse over texts/twitter from [redacted]! it's not fair! I'm sad, i'm only helping her!

These ten tweets that are listed above corroborate the research method that was used and combated for this research. For instance, you can see how each emotional

term correlates with the unigram and each tweet showcase cyberstalking indicative content. Likewise, as stated previously, the two terms “bad” and “sad” were shown within 100 percent of the clustering results and the term “hate” was only seen within 60 percent of the results. Additionally, looking at these ten tweets that are provided, each emotional term is mentioned frequently, for example: “bad” is seen in three out of the ten tweets above with the unigrams: stalking, creep/creepy, and fear attached to the term. “Sad” is seen in four out of the ten tweets above with the unigrams: stalking, creep/creepy, and abuse attached to the term. Lastly, the term “hate” is seen in four out of the ten tweets above and with the unigrams: stalking, technology, and creep/creeper. For instance, with the above fifteen tweets from the preliminary data set (5,000 tweets). “Sad” was in four out of the fifteen tweets with these unigrams: follow, relationship, creep/creepy, technology, abuse, and stalking. “Hate” was in five out of the fifteen tweets with these unigrams: scared, stalker, creep/creepy, messaging, fear, and annoying. Lastly, “bad” was in six out of the fifteen tweets with these unigrams: stalking, violent, harassment, scared, and unwanted. Comparing these two results together it correlates that there is in fact cyberstalking indicative content on Twitter.

The unigrams that the researcher suggested and used within the preliminary data set in fact corroborate that detection of cyberstalking on Twitter is admissible with their use. Furthermore, the emotional terms: “bad”, “hate”, and “sad” beside the unigrams are attached to the tweets mentioned within this chapter from both data sets, proving that these emotional terms and unigrams go hand in hand with regards to cyberstalking indicative content on Twitter. Together with, with the results of the random sample data set, the advanced search, along with k means verifies that all

these findings linked together showcase that this research has sustained true to its objectives and aims.

5.4 Summary

In conclusion to summarise, this chapter describes the clustering algorithm that was used to analyse the data sets required to address the research questions and to test the hypothesised relationships developed in this study, cyberstalking indicative content on social media platforms such as Twitter. The chapter begins with a discussion clustering algorithm, followed using social media with Twitter, the reoccurring themes, and key words within both data sets, fifteen tweets from the preliminary data set (5000 tweets) are provided as examples for the reoccurring themes and emotional terms, and lastly the results from the clustering algorithm along with the advanced search.

The chapter then continues with descriptions of how the clustering algorithm was used on the data that was collected, the clustering algorithm results and findings/measurements, and lastly the emotional terms preferred for this research. Next, the reasoning as to the choice of algorithm and the analysis of the results are discussed. In addition, the limitations and recommendations for future studies are suggested and mentioned, as well as what are expected to achieve and potentially continue to do after this study is concluded. Finally, the next step is to explore the theoretical model that is being proposed using the algorithm, unigrams, and emotional terms, thus focusing on the metrics of cyberstalking.

These three emotional terms are reoccurring through the random sample data set. Another important finding is how these emotional terms are used in tweets on the social media platform Twitter that help showcase cyberstalking indicative content. It is important to understand and realise that this type of study has not been done before

within academia. Therefore, all the findings or results are impactive to the cause and narration of cyberstalking on Twitter. This research sets the tone for this type of work within academia, the narrative within academia towards cyberstalking on Twitter is very vague. This research will help open that conversation and shed light on cyberstalking indicative content on Twitter, dating mining with tweets in correlation to cyberstalking, use of k means algorithm on random sample data sets, unigrams and the emotional terms being used, and so much more. All these methods and results consequently prove, the findings in this chapter help corroborate this research question being asked: *"Which data-mining algorithm is better suited for identifying and detecting cyberstalking on social media platforms?"*

Chapter 6: Development of K Nearest Neighbour Model to Preform Clustering Analysis

6.1. Introduction

This chapter is outlined for the purpose of the K-nearest neighbour (KNN) algorithm. Within this chapter which presents the use of KNN algorithm (K Nearest Neighbour) on the results of the clustering algorithm on the random sample dataset used within Chapter 5. KNN algorithm is generally one of the most known and used data mining or machine learning method. There are various extensions of the use of KNN to be proposed within literature and academia. Although unlike K means, KNN is a Supervised Non-Liner Classification Learning algorithm that uses labelled input dataset to predict the output of the data points. As opposed to the clustering in pattern recognition and machine learning, k-means algorithm and its extensions are always influenced by initializations with a necessary number of clusters. In addition, within this chapter another prime focus is on the results from the KNN model that will be used on the dataset and its variables provided, as well as the liking between those results with the notation of cyberstalking indicative content on Twitter, the possible security metrics, and towards the overall topic as well as future work.

KNN was born out of research done for the armed forces. Fix and Hodge who were two officers of United States Air Force (USAF) School of Aviation Medicine. They both wrote a technical report in 1951 introducing the KNN algorithm for its inclusive purpose. In addition, comparing KNN to the algorithm used in the previous chapter, K-Means: K-means clustering represents and unsupervised algorithm, mainly used for cluster, while as mentioned previously, KNN is a supervised learning algorithm used for classification. For example, KNN uses labelled input dataset to predict the output

of the data points. KNN can and is most used for classification. Moreover, the output is a class membership which predicts a class and a discrete value. An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours.

However, it is noted as one the most unpretentious machine learning algorithms and it can be easily implemented for a varied set of problems, KNN is mainly based on feature similarity. K-Nearest Neighbour or KNN non-parametric algorithm for example, it does not make any assumption about underlying data or its distribution. It is one of the widely used algorithm which depends on its k value (Neighbours) and finds it's applications in many industries like the finance industry, healthcare industry, business industry, banking industry, and many more. Lastly, classic algorithms of datamining, K-means and KNN are used in many applications to exploit data value and enhance the utility of data services.

Currently, the exponential growth of generation of textual documents and the emergent need to structure them increases the attention to the automated classification of documents into predefined categories. There is an inclusive range of supervised learning algorithms that deal with text classification. However, KNN is commonly used as a machine learning tool for the classification of textual documents. Moreover, by again the Internet has abled the fast development of the sharp growth of number of electronic documents. According to Moldagulova and Sulaiman (2017), permitting to experts now about 70 per cent of the digital information which is saved up and used by society is in an unstructured (text) form and only 30 per cent make other types of data. The increase in number of unstructured data exponential eventually led in essence to collapse of traditional system of receiving and distribution of text information, turned routine operation of search and the analysis of necessary

data into the labour-intensive and ineffective process causing information overload of users.

In this situation special relevance is gained by works on creation of systems of processing of text information as even highly skilled experts' experience difficulties on the organisation of search of documents and distribution of the obtained text data on subjects. (Liebowitz and Taylor, 1997). Documents today have an increasing variety of uses, and because of the contribution of the computer in both production and analysis, may be combatted in countless methods. Evermore so, with the rapid increase or growth of the Internet, the rise in unstructured data has renewed and intensified the interest in document classification and text mining (Sebastiani, 2002). Tran, Moon, Le, and Thoma (2001), described the process of classifying medical articles from online journals using constraint satisfaction method. Likewise, Amato, Boselli, Cesarini, Mercorio, Mezzanzanica, Moscato, and Picariello (2015) applied and compared different techniques, in particular explicit rules, machine learning, and linear discriminant analysis based on methods to classify a real time data collection of Web employment offers gathered from various heterogeneous sources with standard job classifications system.

Additional learning systems have been in exploration for managing text classification. Another interest in text classification lays in document representation. For instance, a semantic net for document representation in a five-dimensional space was proposed by M. Lipshutz and S. Liebowitz Taylor (1997). The main idea of their document classification system was based on decomposition of documents into physical, logical, functional components, topical organization, and document class. They developed the classifier which clusters documents by type such as newspapers, business letters, and technical journals.

However, despite the many approaches to solve classification problems the dominant approach is machine learning technique. The advantages of machine learning systems are the efficiency, accuracy, performance, and usability to different domains (Sebastiani, 2002). The main aspects of machine learning paradigm issues such as document representation, classifier construction, and classifier evaluation were surveyed by Sebastiani in 2002. Nowadays classifiers constructed using machine learning tools achieve impressive levels of efficiency and accuracy, bringing to automated classification high quality comparable to manual classification. Many forms of research of recent years prove that this tendency has been seen more and more as the years progress.

Among the existing text classification methods that are emphasized in the previous works are K-Nearest Neighbours (KNN), and other methods such as: Naïve Bayes and Term Graph Model. The comparison of Naïve Bayes, Term Graph and KNN for Text and Document classification made a conclusion that KNN method shows the high accuracy as compared to the Naïve Bayes and Term- Graph algorithms (Bijalwan, Kumar, Kumari, and Pascual, 2014). Although, KNN has a disadvantage that its performance is slow, it is widely used in text classification due to fully dependence on every sample in the training set (Hassanat, Abbadi, and Alhassanat, 2014).

Therefore, in this chapter, the tweets that were extracted, refined, analysed, and visualized for representation are refined even further with the results from K means. Brings forth what this chapter focuses on the results from the first algorithm used K means. For a refresher, K means was used to gather results on the tweets obtain within the five datasets. The results concluded of three emotional terms that are

correlated towards or with cyberstalking on Twitter. A goal for this research is to visualize the cyberstalking tweets in a particular area, on Twitter. While also visualising the clustered emotional terms according to the linking or connection to cyberstalking indicative content on Twitter. However, the researcher decided to have a different look with another algorithm, to compare the results to the previous chapter. As well as, mentioned or stated before, many of the various packages and libraries are provided by R for extracting and processing the data and used for the visualisation of and interpretation of the data. Within this chapter, the researcher is concentrating on the results of the k means clustering algorithm, used from the random sample dataset. Also, again, R language is used for acquisition, pre-processing, analysing and visualization of the twitter data specified. Thus, taking the k means results and extracted then analysed, pre-processed, and now will be examined even further with KNN.

6.2. KNN or K Nearest Neighbour

In present day scenario, machine learning and artificial intelligence are replacing all the conventional computational techniques and programming languages, most importantly machine learning gives computers the ability to learn without being explicitly programmed. The KNN (K Nearest Neighbours) algorithm is a non-parametric, or an instance-based, or seen as a lazy method, and has been regarded as one of the effective methods in data mining and machine learning (Zhang, Liu, Hu, Lv, Gong, Sha, and Wu, 2017). The principle of KNN algorithm is that the most similar samples belonging to the same class have high probability. Generally, the KNN algorithm first finds k nearest neighbours of a query in training dataset, and then predicts the query with the major class in the k nearest neighbours. Therefore, it has recently been selected as one of top 10 algorithms in data mining (Wu, Kumar,

Quinlan, Ghosh, Yang, Motoda, McLachian, Liu, Philip, and Zhou, 2008). As well known, KNN algorithm is often sensitive to the selection of the k value. Although efforts have been focused on this topic for a long time, setting k value is still very challengeable in KNN algorithm (Zhang, WU, and Zhu, 2010). Moreover, it has been proved that a fixed k value is not suitable for many test samples in each training K-nearest neighbour (KNN) classification method originally developed in the probabilistic framework that has serious difficulties to correctly classify the close data points (objects) originating from different classes.

The *K* Nearest Neighbour method has widely been used in the applications of data mining and machine learning due to its simple implementation and distinguished performance. However, setting all test data with the same *k* value in the previous KNN methods has been proven to make these methods impractical in real applications. However, to apply KNN the researcher needs to choose an appropriate value for k, and the success of classification is very much dependent on this value. Moreover, more details on how the researcher chose which value to use for k and why that value is better suited for this study is further on within this chapter. Additionally, looking at the method that is KNN in a sense, the KNN method is biased by k. There are many ways of choosing the k value, but a simple one is to run the algorithm many times with different k values and choose the one with the best performance. Consequently, why three different forms of the dataset were used for this section and for this algorithm on the present study.

Furthermore, it is important to know that there are various modifications of KNN algorithm. It was shown that a flexible K-Nearest Neighbours algorithm with combination of K-variable algorithm and weighting algorithm enhances the efficiency of text classification (Yunliang, Lijun, Xiaodong, and Quan, 2009). Another

modification of KNN algorithm is a combination of eager learning with KNN classification (Tao Dong and Cheng, 2012), which improved the efficiency and increased the accuracy of classification. A novel KNN classification algorithm combining model and evidence theory helps to overcome the shortage of lazy learning in traditional KNN method such as time-consuming (Guo, Ping, Chen, 2006). Based on the discussion above, it shows that many researchers have tried various methods to combine different classification approaches to increase the classification accuracy and reduce time consumption. Despite the KNN algorithm is easy to use and effective in general, the performance of KNN algorithm depends mostly on the allocation of the training set. Considering that the textual data distribution is uneven, it was proposed to use a modified KNN algorithm based in integration the density of the test sample and the density of its nearest neighbours (Li and Chen, 2011 & Shi, Li, Liu, Zhang, and Song, 2011). This is to decrease the effect of the uneven data distribution to the classification they amplify the distance between the test sample and samples in the sparse area and reduce the distance between the test sample and samples in the dense area. To solve the problem of the uneven distribution of training samples, it was presented an algorithm based on clustering the training samples making a relatively uniform distribution of training samples (Zhou, and Wang, 2010).

Furthermore, the K-nearest neighbour (KNN) method is a well-known classification algorithm used in pattern recognition. In the original voting KNN, the object is assigned to the majority class according to its K nearest neighbours (KNNs), and the distances between the object and its neighbours are ignored. The KNN or k-nearest neighbour algorithm is one of the effective machine learning algorithms and is an example of instance-based learning, where new data are classified based on

stored, labelled instances. In the KNN algorithm, K specifies the number of neighbours, and its algorithm is as follows:

- Choose the number K of neighbour.
- Take the K Nearest Neighbour of unknown data point according to distance.
- Among the K-neighbours, Count the number of data points in each category.
- Assign the new data point to a category, where you counted the most neighbours.

For the Nearest Neighbour classifier, the distance between two points is expressed in the form of Euclidean Distance. For instance, in mathematics, the Euclidean Distance between two points in Euclidean space is the length of a line segment between the two points. It can be calculated from the Cartesian coordinates of the points using the Pythagorean theorem, therefore occasionally being called the Pythagorean distance. Moreover, Euclidean Distance: is the most widely used one as it is the default metric that SKlearn library of R Programming and another code (Python) uses for K-Nearest Neighbour. It is a measure of the true straight-line distance between two points in Euclidean space.

More specifically, the distance between the stored data and the new instance is calculated by means of a similarity measure. This similarity measure is typically expressed by a distance measure such as the Euclidean distance, cosine similarity or the Manhattan distance. In other words, the similarity to the data that was already in the system is calculated for any new data point that is input into the system. After, use this similarity value to perform predictive modelling. Predictive modelling is either classification, assigning a label or a class to the new instance, or regression, assigning

a value to the new instance. Whether it is classified or assigned a value to the new instance depends of course on how a person composes his or her model with KNN. The k-nearest neighbour algorithm adds to this basic algorithm that after the distance of the new point to all stored data points has been calculated, the distance values are sorted, and the k-nearest neighbours are determined. The labels of these neighbours are gathered, and a majority vote or weighted vote is used for classification or regression purposes. Moreover, the higher the score for a certain data point that was already stored, the more likely that the new instance will receive the same classification as that of the neighbour. In the case of regression, the value that will be assigned to the new data point is the mean of its *k* nearest neighbours.

Nowadays, privacy protection has become an important issue with data mining. For instance, K-means clustering and KNN classification are two very popular data mining algorithms, which have been widely studied in the past decade (Zhao, Hu, Xiong, Tianm Xiang, Zhou, and Li, 2021). As classic algorithms of data mining such as K-means and KNN are used in many applications to exploit data value and enhance the utility of data services. As the same time, these methods are originally designed for analysing plain-text data and have not taken privacy protection into account. However, many other works have taken and or studied the approaches for privacy preserving K-means clustering and KNN classification for instance (Doganay, Pedersen, Saygin, Savas, and Levi, 2008) which gives great insight on the matter.

In KNN algorithm, to measure a document relevancy to a given query again is the Euclidean distance between the query vector and the document vector. This metrics is modestly successful. The ways of improving on it are described by Lars Elden (2007), that shows of replacement of the term document matrix by a low-rank approximation in an attempt to capture the important information and discard the

irrelevant details (Elden, 2007). It was found that KNN shows the best result with accuracy among Naïve Bayes, Term Graph and KNN algorithms for Text and Document classification tasks (Bijalwan et al., 2014). However, comparison of Naive Bayes, K-Nearest Neighbours and Support Vector Machine classification methods applied to predict user's personality based on texts written on Twitter shows that Naive Bayes method performs better than the other methods (Pratama and Sarno, 2015). This is because Naive Bayes uses pure probability calculations on existing features. In certain cases, the KNN algorithm shows a lower speed and applicability to text categorization because of high dimension of text vectors (Yan, 2010). This indicate that KNN requires more time for classifying documents when a large number of training examples are given. Although, in this study the training and test datasets were small enough with regards to the results from K-means, for KNN to work more than efficiently for the purpose of this research.

6.2.1 How KNN is Used

A supervised machine learning algorithm like KNN (as opposed to an unsupervised machine learning algorithm like K-means) is one that relies on labelled input data to learn a function that produces an appropriate output when given new unlabelled data. An unsupervised machine learning algorithm makes use of input data without any labels in other words, no teacher (label) telling the child (computer) when it is right or when it has made a mistake so that it can self-correct. Unlike supervised learning that tries to learn a function that will allow us to make predictions given some new unlabelled data, unsupervised learning tries to learn the basic structure of the data to give more people insight into the data.

A classification problem has a discrete value as its output. A great example for a classification problem is the pizza example, “likes pineapple on pizza” and “does not like pineapple on pizza” both are discrete examples, there is no middle ground for each, either you like pineapple on pizza or you do not like pineapple on pizza. Furthermore, a regression problem has a real number (a number with a decimal point) as its output. For example, we could use the data in the table below to estimate someone’s weight given their height. The k-nearest neighbours (KNN) algorithm is a supervised machine learning algorithm that can be used to solve both classification and regression problems. It’s easy to implement and understand but has a major drawback of becoming significantly slow as the size of that data in use grows. Both classification and regression are talked about in greater detail below.

6.2.1a. Classification

There are two types of KNN algorithms that can be used, one is classification, and the other is regression. KNN classification algorithm first selects k closest samples (i.e., k nearest neighbours) for a test sample from all the training samples, and then predicts the test sample with a simple classifier, e.g., majority classification rule. Liu, Zhang, Zhao, and Mo, (2010), designed a new anomaly removal algorithm under the framework of KNN classification (Liu et al.,). The framework which adopts mutual nearest neighbours whose advantage is that pseudo nearest neighbours can be recognised instead of k nearest neighbours to determine the class labels of unknown samples, which is beneficial for any study. Weinberger and Saul, (2009) used semi-definite programming to learn a Mahanalobis distance metric for KNN classification and implemented the target that k nearest neighbours always belong to the same class to optimise the measure metric, which then samples from different classes are separated by a large margin (Weinberger et al.,). KNN classifier needs to keep all the

training examples in memory, to search for all the *K* nearest neighbours for a test sample. Despite of the KNN classifier for imbalanced data (Zhang, 2010) there is no work on constructing KNN classifier in the field of cost-sensitive learning (called as CSL). The main challenge is how to make the KNN classifier sensitive to the topic at hand for the classification, because the *K*-nearest neighbours is only a small subset of the whole training sample space.

There are two main research directions in any case. One is to set a proper K value as mentioned. Another is the distance function for identifying K nearest neighbours. For setting K value, a usually used method is the cross validation in probability theory. It is useful for identifying a proper K value when a training dataset is given. However, training samples are distributed with different densities in the training sample space. This raises a new challenging issue that different samples need different K values for class prediction. Recently, suggested by: Cheng, Zhang, Deng, Zhu, Zong (2014) studied the computation of parameter K for KNN classification, which is an optimal value for each new data. Zhang, Li, Zong, Zhu, and Wang (2017) mentions, designed a KNN algorithm with data driven K parameter computation. Brought forward by, Zhang, Li, Zong, Zhu, and Cheng (2017) who designed an algorithm to efficiently learn K for KNN Classification.

Moreover, Goldberger, Roweis, Hinton, and Salakhutdinov, (2004), proposed a novel non-parametric KNN classification that learns a new quadratic distance metric and calls neighbourhood component analysis (NCA) method (Goldberger et al.,). Moreover, this method focuses on the learned distance to be lowrank, by saving the storage and search costs for studies. Jamshidi and Kaburlasos (2014) proposed an effective synergy of the Intervals' Number K -Nearest Neighbour classifier, and the gravitational search algorithm (GSA) for stochastic search and optimisation (Jamshidi

et al.,). Also, Saini, Singh, and Khosla, (2013), presented an application of K -Nearest Neighbour (KNN) algorithm as a classifier for detection of QRS-complex in ECG (Saini et al.,). In addition, this algorithm uses a digital band pass filter to reduce the interference present in ECG false detection signal. For avoiding the influence of k value, Varmuza, Filzmoser, Hilchenbach, Krüger, and Silen, (2014), used the repeated double cross validation method to search an optimum k for k nearest neighbour classification (Varmuza et al.,2014).

6.2.1b. Regression

The KNN regression has been widely used and studied for many years in pattern recognition and data mining. In pervious regression analysis, Burba, Ferraty, and Vieu, (2009), utilised kernel estimator based some asymptotic properties of the KNN to improve the performance of KNN regression (Burba et al.,). Moreover, the purpose of their work utilised local adaptive bandwidth to study the non-parametric KNN algorithm. Ferraty and Vieu (2006), utilised the functional version of the Nadaraya–Watson kernel type estimator to construct the non-parametric characteristics of KNN algorithm for estimation, classification, and discrimination on high dimensional data (Ferraty et al.,). Additionally, in the theory of KNN algorithm, Mack (1981), studied the L 2 convergence and the asymptotic distribution and Devroye (1981) proved the strong consistency and the uniform convergence of k NN algorithm (Devroye 1981). Likewise, Hu, Jain, Zhang, Schmidt, Gomadam, and Gorka, (2014), proposed a data-driven method for the battery capacity estimation and used a non-linear kernel regression model based on the KNN to capture the dependency of the capacity on the features. Furthermore, this work also utilises the adaptation of particle swarm optimisations to find the feature weights for the KNN regression model (Hu et al.,). Goyal, Chandra, Singh, (2014), took the interrelatedness of these metrics into

account and statistically established the extent to improve the explanatory power of multiple linear regression.

They continued to conduct a stepwise regression to identify influential metrics to avoid over fitting of data and proposes suitability of KNN regression in the development of fault prediction model (Goyal et al.,). Cycle time of wafer lots for semiconductor fab was a critical task, therefore, Nia, Qiao, Li, and Wu, (2021), combined the particle swarm optimization with a Gaussian mutation operator and a simulated weight of the features for KNN regression, and then used it to predict the cycle time of wafer fab (Ni et al.,). Zhou (2005) proposed semi supervised regression with co-training (Zhou and Li, 2005), which employed two KNN regressors with different distance metrics, each of which labelled the unlabelled data for the others during the learning process.

6.3. KNN in Exercise

Likewise, KNN is an algorithm that is used multiple ways as stated previously. Here is a prime example of KNN. Let's assume an individual has several groups of labelled samples. The items present in the groups are homogenous in nature. Now, suppose he or she has an unlabelled example which needs to be classified into one of the several labelled groups. How is that done exactly, well using KNN. K nearest neighbour is a effective algorithm that stores all available cases and classifiers new cases by a majority vote of its k neighbours. This algorithm segregates unlabelled data points into well-defined groups. Another way of looking at KNN, the k-nearest neighbour algorithm accumulates all the available data and classifies a new data point based on the similarity measure (used for distance functions). This means when new data appears it then can be easily classified into a well-suited category by using KNN algorithm.

Choosing the number of nearest neighbours i.e., determining the value of k plays a significant role in determining the efficacy of the model. Thus, selection of k will determine how well the data can be utilised to generalise the results of the KNN algorithm. A large value has benefits which include reducing the variance due to the noisy data; the side effect being developed a bias due to which the learner tends to ignore the smaller patterns which may have useful insights. In addition, here are some pros and cons to KNN. Pros: the algorithm is highly unbiased in nature and makes no prior assumption of the underlying data. Being simple and effective in nature, it is easy to implement and has gained good popularity. Now on to the cons: indeed, it is simple, but KNN algorithm has drawn a lot of flak for being extremely simple. If people take a deeper look, this does not create a model since there is no abstraction process involved. Yes, the training process is fast as the data is stored verbatim (hence known as a lazy learner), but the prediction time is high with useful perceptions missing at times. Therefore, building this algorithm requires time to be invested in data preparation (especially treating the missing data and categorical features) to obtain a robust model.

One of the most essential aspects while using KNN, is data collection, now the data that was used for this chapter was the data collected from previous chapters within this thesis and then used to analysed within the previous chapter using K-means clustering algorithm. As a reminder, the data that was used for KNN in this chapter, was the clustering results from the previous chapter and they consist of the three emotional terms: “bad, “sad”, and “hate”. Along, with their values rendered from the clustering algorithm K means. However, it is important to note, when having the data imported that is being used and make an observation on the variables that are labelled within RStudio. Hence, if they are numeric, a character, integer, or factor value in R

programming. The reason as to why the above statement is important is because, KNN can only be run on a dataset that is either numeric, factor, or integer and cannot be run on any datasets that are character or text based. Therefore, why the researcher used the K-means clustering results from the previous chapter and not the random sample dataset. Mainly because, the random sample dataset is formed on text-based data were as the result from the clustering algorithm, k means is not. Hence, why this part of the research method is done this way since the K-means clustering results are numeric and the random sample dataset is text based. However, there is a way to rewrite the code and change the dataset, but within the time frame for this study and PhD the researcher did not have time to do that and used the prior results from chapter 5 instead.

6.4. Using KNN with K-Means Results

For instance, as the researcher previously stated he or she used the results from K-means. In these results there were three emotional terms: “bad’, “sad”, and “hate” seen in Chapter 5. With each term it was a value within the cluster they (the emotional term) were represented in within each dataset. Then, all five datasets had at least 2-3 emotional terms and their respectful values or clusters. After these results, the researcher came to the decision of using KNN for the final algorithm in this study. Now, KNN is used to analysis these results even further from the previous chapter and algorithm used. How the accuracy within these terms differ or even correlate, as well as how the RMSE which is: The RMSE corresponds to the square root of the average difference between the observed known outcome values and the predicted values, RMSE = mean (observed(s) – predicted(s))^2) %>% sqrt(). The reasoning as to why this was done was for the purpose see the repeated cross-validation is from the K-means results from *Chapter 5: Clustering Algorithm.*

However, as previously mentioned KNN is used on numeric, factor, or integer datasets and not character or text-based datasets. Therefore, KNN was used for the simplistic reason that the algorithm is used for numeric, factor, or integer-based datasets. Therefore, using the clustering results within KNN for a comparison look on the algorithm performance. Moreover, the results from the previous chapter were used and taken and imported into RStudio for KNN to analyse. As for importing the K-means results into RStudio, it is as straight forward as before. Each result was saved on the researcher's computer and then imported via csv file into RStudio. Conversely, once that was completed the researcher had to make some changes to the format of the values while in RStudio, but that will be detailed below.

6.4.1 Importing the K-means Results

As stated above, the import process is seemingly straight forward. However, an important tip is to make sure all the libraries that are being used for KNN in RStudio are either installed or already on the system. As previously mentioned within R the libraries and its packages are all available within the system and there are vast majorities to choose from. For instance, the libraries that were used for within this section are: library(caret), library(pROC), and library(mlbench). Library(caret) is used for: the caret package (short for Classification and regression Training) contains functions to streamline the model training process for complex regression and classification problems. The (Caret) package or library loads packages as needed and assumes that they are installed. However, if a modelling package is missing, there is a prompt to install it. Library(pROC) is used for: as a tool for visualizing, smoothing, and comparing receiver operating characteristic (ROC curves). (Partial) area under the curve (AUC) can be compared with statistical tests based on U-statistics or bootstrap. Confidence intervals can be computed for (p)AUC or ROC curves. Lastly,

library(mlbench) is used for: MLBench is a framework for distributed machine learning. Its purpose is to improve transparency, reproducibility, robustness, and to provide fair performance measures as well as reference implementations, helping adoption of distributed machine learning methods both in industry and in the academic community. Once these files are installed within RStudio the next steps are for formatting the dataset being used and for beginning uses of KNN.

In addition, each result was saved in a csv file all together and separately on the researcher's computer. Respectively, as well all the files contain the emotional term and all its values from the clustering algorithm. Once the files are located within the computer that is being used, they are imported into RStudio. Once the files are imported, double check the value of each component to see if they need to be changed or formatted differently in order to conduct the KNN algorithm.

Below shows the data being imported into R and how the data is made up. The term and hashtag represented below are how the researcher broke up his or her code to make it seamless and explain what he or she was doing.

#Classification

- data <- read.csv(file="~/Desktop/knn_terms.csv",header= T)
- str(data)
- data$emtional.term[data$emtional.term == "bad"] <- 1
- data$emtional.term[data$emtional.term == "sad"] <- 2
- data$emtional.term[data$emtional.term == "hate"] <-3
- data$emtional.term<- factor(data$emtional.term)
- str(data)

Here are the classification steps within the RStudio console that the researcher used for the preparation for KNN. The first step was done for the purpose of importing the dataset, again are the results containing the emotional terms and their respected clusters from the previous algorithm and which was used for the KNN algorithm. The second was to see the structure of the full dataset imported in R and to see if anything needed to be changed. Once the above is completed the model conduction and its performance will take place on the emotional terms dataset.

Moreover, when having a look at the string str(data) people can see the data.frame which is the dataset made of 13 observations and 3 variables. Which are the values of the clusters and the three emotional terms from the K-means results. Followed by the dataset which has various numbers from 1-5 which make up each dataset that the emotional terms come from, for instance the original 5 csv files from the random sample dataset. Followed by the emotional terms “bad”, “sad”, and “hate” and lastly the cluster analysis.

- str(data)
    - 'data.frame': 13 obs. of 3 variables:
    - $ dataset : int 3 2 1 4 5 1 2 3 4 5 ...
    - $ emtional.term : chr "bad" "bad" "bad" "bad" ...
    - $ cluster.anaylsis: chr "0.04028436 , 0.03074485 , 0.02847844 , 0.01930147, 0.02119701 , 0.0540153 , 0.0247795 , 0.02516109 , 0.02492558"| __truncated__ "0.40700902\n0.01912046\n0.02410901\n0\n0.0125969\n0.0401733\n0.02637244\n0.03777879\n0.01763409\n0.04125874\n0."| __truncated__

"0.02881237\n0.02439673\n0.02258292\n0.01996528\n0.02882096\n0.02
15311\n0.01606805\n0.02865762\n0.0252646\n0.026"| __truncated__
"0.02593965\n0.02626094\n0.03513174\n0.02550091\n0.01535166\n0.50
215916\n0.03354037\n0.02945253\n0.01840491\n0.0"| __truncated__ ...

The next three bullet points or steps are very important. This procedure was done for the purpose of using the variable emotional.term as a factor, therefore taking the variable emotional.term for each emotional word (bad, sad, and hate) as seen below and making them correlate with the numbers 1,2,3 meaning bad=1, sad=2, and finally hate=3, oppose to the characters used for counting the term. The reasoning as to why this was done because the default setting in R as the variable labelled as a character vector and for the purpose of this study the variable needed to be a ranking numeric or factor for keeping the tallies for each time the emotional term rather than the overall usage for each emotional term. Thus, labelling those variables as numeric/factor: 1,2, and 3 and not having them as character variables or having the overall tally of each variable separately.

- data$emtional.term[data$emtional.term == "bad"] <- 1
- data$emtional.term[data$emtional.term == "sad"] <- 2
- data$emtional.term[data$emtional.term == "hate"] <-3

Then lastly the following stringer was preformed to make the emotional terms variable as a factor (data$emtional.term<- factor(data$emtional.term), for the purpose of this research and the type of dataset, which is better suited for the KNN algorithm compared to how it was originally a character variable. Again, str(data) is used to just make sure all the correct changes were made appropriately and to verify that the process can continue.

- data$emtional.term<- factor(data$emtional.term)
- str(data)

6.4.1a. Data Partition

After following the above steps or guidance, the data partition is underway.

- # Data Partition
- set.seed(1234)
- ind <- sample(2, nrow(data), replace= T, prob = c(0.7, 0.3))
- training <- data[ind == 1,]
- test <- data[ind== 2,]

The set.seed is used to have a set training and test model for KNN to use. In this case, set.seed was used at (1234) while then an independent sample is used with size 2 and number of rows which is (data). However, it is important to understand that the set.seed can be fixed to any variable as well as the independent sample size. This is a sampling with replacement so replace=T (true) with a probability of 70 per cent for training data and 30 per cent for test data. Now, with the training the rows in the data with regards to independent sample to: == [1,] with 1 and a comma with nothing after it meaning all the columns are included in this process, which is what is needed for this study, however, that can be different compared to the dataset and study being worked on. Once, these two strings are run through RStudio, as can be seen from the Figures shown below, the first chart is the training data: which consists of 12 observations and 3 variables. Followed by the test data: which contains the following 1 observation and again 3 variables. The way the training and test data are set up this

way is because, with the fact that the training data was 70 per cent and the test data was the other 30 per cent. The researcher did try and have the percentage as 60 and 40, however the results were inconclusive therefore the reasoning for having the test data the training data set to 70/30.

		dataset	emtional.term	cluster.anaylsis
1	3	1	0.04028436 , 0.03074485 , 0.02847844 , 0.01930147, 0.02119701 , 0.0540153 , 0.0247795 , 0.02516109 , 0.02492558 , 0.02074288 , 0.03389831 , 0.01943844	
2	2	1	0.40700902 0.01912046 0.02410901 0 0.0125969 0.0401733 0.02637244 0.03777879 0.01763409 0.04125874 0.03481625 0.01346046	
3	1	1	0.02881237 0.02439673 0.02258292 0.01996528 0.02882096 0.0215311 0.01606805 0.02865762 0.0252646 0.02657219 0.03381643 0.01811594	
4	4	1	0.02593965 0.02626094 0.03513174 0.02550091 0.01535166 0.50215916 0.03354037 0.02945253 0.01840491 0.02384063 0.01639344 0	
6	1	2	0.01777251 0.03042789 0.03498779 0.02757353 0.01745636 0.05736138 0.02183956 0.02117214 0.03066385 0.02556681 0.0273794 0.0237581	
7	2	2	0.40700902 0.01912046 0.02410901 0 0.0125969 0.0401733 0.02637244 0.03777879 0.01763409 0.04125874 0.03481625 0.01346046	
8	3	2	0.01777251 0.03042789 0.03498779 0.02757353 0.01745636 0.05736138 0.02183956 0.02117214 0.03066385 0.02556681 1.03887689 0.010799136	
9	4	2	0.035468502 0.028345144 0.008782936 0.034001214 0.043912888 0.032078964 0.027950311 0.066528067 0.040899796 0.048007838 0.013583138 0.036480532	
10	5	2	1.037098791 0.028345144 0.008782936 0.034001214 0.043912888 0.032078964 0.02542726 0.066528067 0.040899796 0.048007838 0.013583138 0.036480532 1.415146568	
11	1	3	0.008293839 0.000633914 0.013018714 0.002757353 0.006234414 0.005258126 0.005039899 0.001534213 0 0.004341534 0.000651891 1.03887689	

		dataset	emtional.term	cluster.anaylsis
12	2	3	0.01943095 0.0290631 0.01572327 0.02184525 0.03391473 0.02402521 0.02529602 0.01729631 0.02130786 0.02132867 0.01740812 0.02075154	
13	3	3	0.008293839 0.000633914 0.013018714 0.002757353 0.006234414 0.005258126 0.005039899 0.001534213 0 0.004341534 0.000651891 1.03887689	

Showing 1 to 12 of 12 entries, 3 total columns

Figure 38. The R Console of the Training Data completed in RStudio

		dataset	emtional.term	cluster.anaylsis
5	5	1	0.02593965 0.02626094 0.03513174 0.02550091 0.01535166 0.50215916 0.03354037 0.01834862 0.01840491 0.02384063 0.01639344 1.035429584 1.776301614	

Showing 1 to 1 of 1 entries, 3 total columns

Figure 39. The R Console of the Test Data completed in RStudio

6.5 KNN Model

KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression). This KNN model that was made and performed on the datasets that was used was a classification model. The final process to have the KNN algorithm used is conducting the KNN model, which is the last step before the performance portion for the KNN algorithm process in RStudio. The functions and strings that are mentioned below are imperative towards the usage of the KNN model that is being used for this study. trControl is named and used for the train control from correct package for developing the model and then use the method “repeatedcv” or repeated cross validation which is used for: in create multi fold, code iterates over multiple times (given by repeats in train Control) syntax in R) for each of the k cross fold (given by number). In cross fold, while using CV, it is a one-time process on each of the fold (set by using numbers in train control). Then the number 10 is used for recent iterations

and the cross validation is set to repeat 3 times; the number of complete set of folds to repeat this validation.

Promptly, before the model is fit, set.seed(222) is used to have repeatability of the outcome and then the model is fit. So, the model is stored in the string (fit) followed by the train function. Train function is used: train can be used to tune models by picking the complexity parameters that are associated with the optimal resampling statistics. By default, the function createGrid is used to define the candidate values of the tuning parameters. Also, if the user or programmer needs to want to change the tuning parameters, the user can also specify their own. The response variable after the train function is admit, therefore the model that is being created is using admit versus all the independent variables. For all the variables there is a (.), data = training, method is then KNN, tuneLength = 20, and lastly the trControl that was used and created earlier. Once, all of these are finished the user can then plot the model and check its performance with fit and input each variable the user wants the KNN model to perform on. For instance, in the example below fit <-train(dataset ~., the dataset called dataset it the main focus. Therefore, this was done to the overall dataset, the emotional terms dataset, and lastly the cluster analysis results from chapter five using the K-means algorithm.

- trControl <- trainControl(method = "repeatedcv",
    - number = 10,
    - repeats = 3)
- set.seed(222)
- fit <- train(dataset ~ .,
    - data =training,
    - method = "knn",

    - tuneLength = 20,
    - trControl = trControl)
- plot(fit)
- fit

Moreover, again it is important to know that with using KNN model each section of the dataset that was being used was done separately. Therefore, the clusters, emotional terms, and the overall dataset was conducted and performed by themselves. Also, each KNN model performance for each counterpart is throughout down below in more detail.

6.6 Model Performance and Results

As stated previously, the KNN model was used on the overall dataset from the clustering algorithm results seen and highlighted in chapter 5. As well as, on the other contributing factors such as the emotional terms results, and the cluster analysis results again from chapter 5. Below are the results from KNN in RStudio, first it has the overall dataset second, it has the emotional terms and lastly, it has the cluster analysis results. In the above section the strings and lines of code were used for each result, but the fit<-train line again changed for each variable thus giving different results catered to the variable being used. Therefore, for the overall dataset it would be fit <-train(dataset ~., data= training, method = “knn”, tuneLength = 20, trControl = trControl). For the emotional terms data set: fit <-train(emotional.term ~., data= training, method = “knn”, tuneLength = 20, trControl = trControl). Finally, for the cluster analysis results: fit <-train(cluster.analysis ~., data= training, method = “knn”, tuneLength = 20, trControl = trControl) then plot and see the results as the model preforms for each dataset.

Dataset Console

> fit
k-Nearest Neighbors

12 samples
2 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 11, 11, 11, 11, 11, 11, ...
Resampling results across tuning parameters:

k	RMSE	Rsquared	MAE
5	1.261916	1	1.254074
7	1.201125	1	1.190387
9	1.201125	1	1.190387
11	1.199764	NaN	1.189697
13	1.199764	NaN	1.189697
15	1.199764	NaN	1.189697
17	1.199764	NaN	1.189697
19	1.199764	NaN	1.189697
21	1.199764	NaN	1.189697
23	1.199764	NaN	1.189697
25	1.199764	NaN	1.189697
27	1.199764	NaN	1.189697
29	1.199764	NaN	1.189697
31	1.199764	NaN	1.189697
33	1.199764	NaN	1.189697
35	1.199764	NaN	1.189697
37	1.199764	NaN	1.189697
39	1.199764	NaN	1.189697
41	1.199764	NaN	1.189697
43	1.199764	NaN	1.189697

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was k = 43.

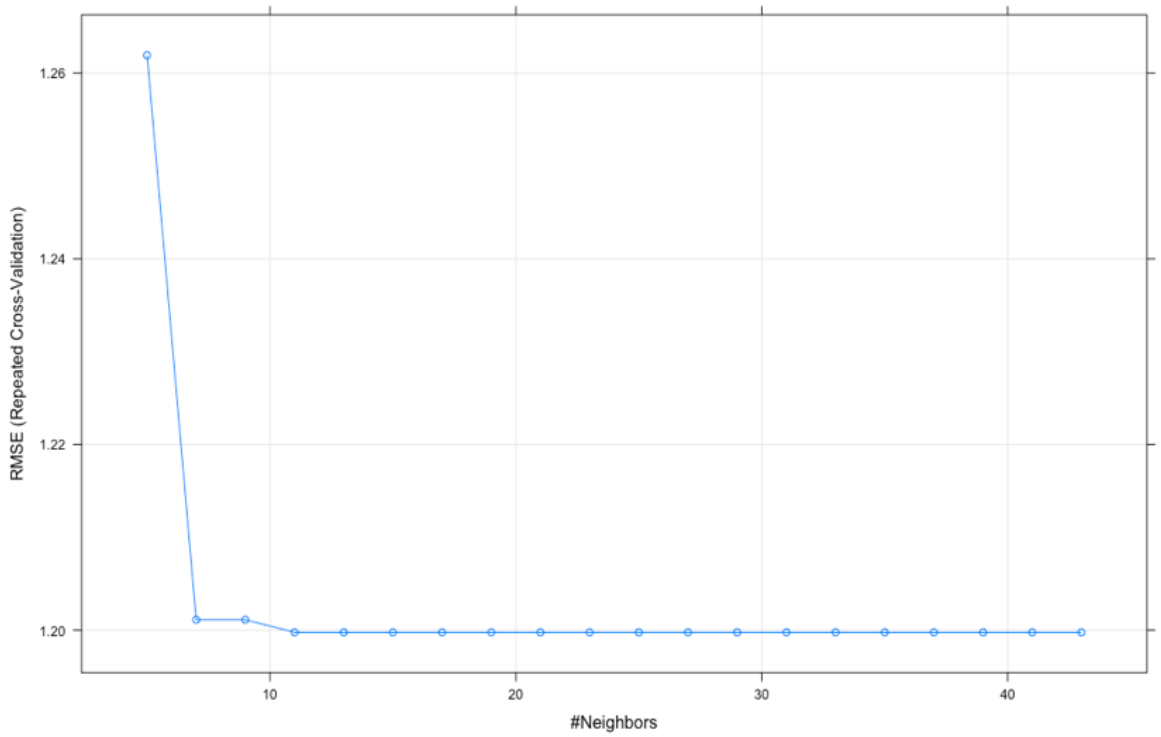


Figure 40. KNN Algorithm Model Performance on the overall dataset results taken from K-Means: RMSE

In the above dataset console for the KNN console results, it is verified that the model that was used was in fact K-Nearest Neighbours. The sample is 12 and the predictor is 2, the resampling: Cross-Validated (10-fold, repeated 3 times), for each cross validation they were split 10 folds or parts, so nine of the are used for creating the model and the final one is used for accessing the model. It yields the: k RMSE and MAE results for each potential k. Along with the summary of sample sizes and the model resampled results across tuning parameters. Lastly, RMSE was used to select the optimal model using the smallest value. The final value used for the model was k = 43. Therefore, while using the KNN model and with its performance and the overall dataset results for K-means the best or optimal K to use for this model was 43, thus the optimal model for k value for the overall dataset used for KNN would be when k = 43 showing these results: 1.199764 and 1.189697. Thus, making the k value 43 the

most frequent label within the overall dataset that was performed in the classification KNN model.

Within Figure 39 which showcases the KNN model for a better understanding or clearer view for the overall dataset plotted in RStudio. As it is shown in the figure the x axes were labelled as the K Neighbours with the point the ending point is 43. Thus, showcases the optimal model using the smallest value for k is 43. Thus, the y axes were labelled as the RMSE, which again is, the correspondence to the square root of the average difference between the observed known outcome values and the predicted values, RMSE = mean (observed(s) – predicted(s))^2) %>% sqrt(). As can be seen having the feature of the line graph in RStudio which plots the KNN model does helps illustrate the KNN numeric results clearer for the viewer to see.

Emotional Terms

```
> fit
k-Nearest Neighbors

12 samples
2 predictor
3 classes: '1', '2', '3'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 11, 10, 10, 11, 10, 10, ...
Resampling results across tuning parameters:
```

k	Accuracy	Kappa
5	0.07971014	-0.07971014
7	0.15942029	0.00000000
9	0.15942029	-0.06060606
11	0.39855072	0.03703704
13	0.31159420	-0.01754386
15	0.35507246	-0.01851852
17	0.28985507	-0.01666667
19	0.26811594	0.01587302
21	0.26811594	0.00000000
23	0.33333333	-0.01754386
25	0.33333333	0.00000000

27	0.28985507	-0.01666667
29	0.28985507	-0.01666667
31	0.24637681	0.00000000
33	0.31159420	-0.01754386
35	0.28985507	0.01666667
37	0.33333333	0.00000000
39	0.26811594	-0.01666667
41	0.33333333	0.00000000
43	0.33333333	0.00000000

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 11.

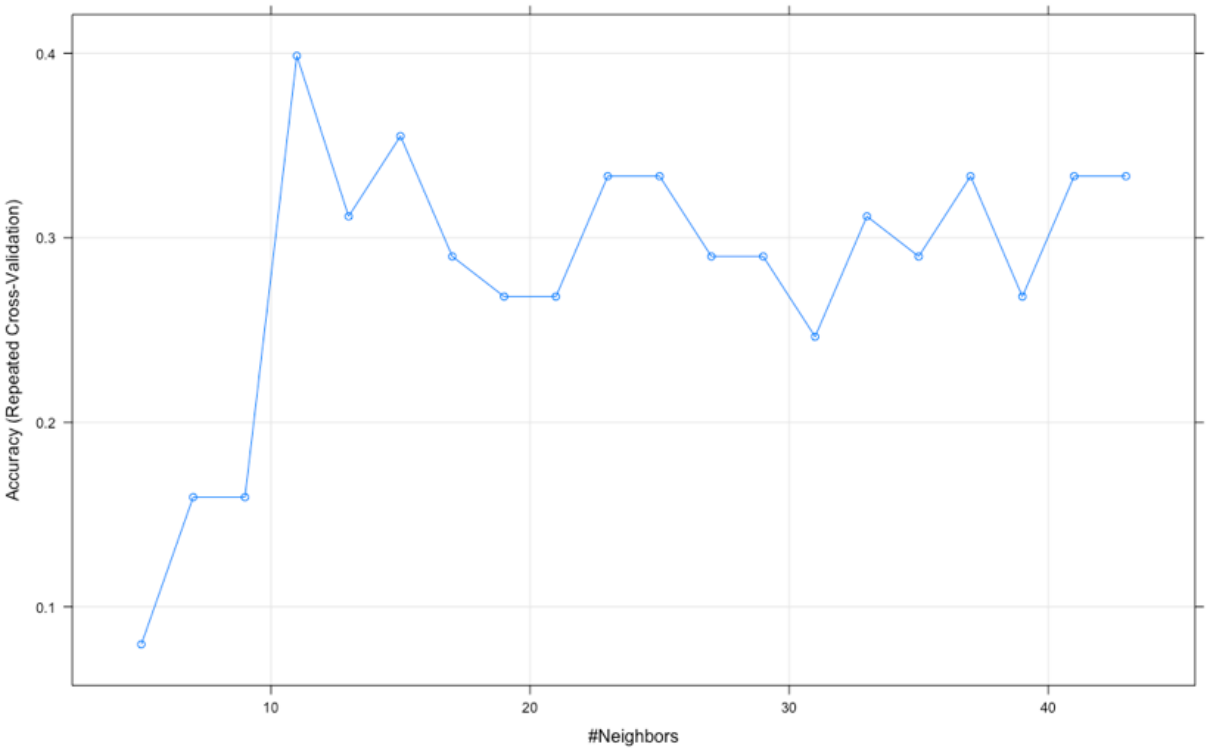


Figure 41. KNN Model Performance: on the Emotional Terms results taken from K-means: ARCV

Overhead, in the above emotional terms console, which is for the KNN console results, and it is verified that the model that was used was in fact K-Nearest Neighbours. Interestingly enough slight similar to the overall dataset, the sample is 12 and the predictor is 2, with 3 classes: '1', '2', '3' these classes are the emotional terms (“bad” = 1, “sad” = 2, and “hate” = 3) as the same as before, the resampling: Cross-Validated (10 fold, repeated 3 times), for each cross validation they were split 10 folds

or parts, so 9 of the are used for creating the model and the final one is used for accessing the model. As can be seen it yields the: k Accuracy and Kappa results for the various values of k. Along with the summary of sample sizes and the model resampled results across tuning parameters. Accuracy was used to select the optimal model using the largest value. The final value used for the model was k = 11. Therefore, while using KNN with the emotional terms results for K-means the best or optimal K to use for this model was 11, the optimal model for KNN for the emotional term dataset that was used would be when k = 11 showing these results: 0.39855072 and 0.03703704. Likewise, as seen above, for the emotional term dataset making the k value 11 the most persistent label within the emotional term dataset that was performed in the classification KNN model.

Moreover, Figure 40 showcases the KNN model for the emotional terms plotted in RStudio. As it is shown in the figure the x axes were labelled as the K Neighbours with the highest point located at 11. Thus, the y axes were labelled as the accuracy, which is the repeated cross validation, again for each cross validation they were split 10 folds or parts, so nine of them are used for creating the model and the final one is used for accessing the model. As the user can see, for all the values of k the results are wavering along the line graph. With the highest point for k being 11 as the KNN model has performed and the lowest being 31. Affectively, RStudio along with KNN have many features like this line graph which plots the KNN model that was used for this study, which does helps illustrate the KNN numeric results clearer for the viewer to see.

Cluster Analysis

```
> fit
k-Nearest Neighbors
```

12 samples
2 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 11, 9, 11, 11, 10, 10, ...
Resampling results across tuning parameters:

k	Accuracy	Kappa
5	0.000	0.00000000
7	0.125	0.08333333
9	0.000	0.00000000
11	0.125	0.08333333
13	0.000	0.00000000
15	0.000	-0.08333333
17	0.125	0.08333333
19	0.000	0.00000000
21	0.000	-0.08333333
23	0.000	-0.08333333
25	0.000	0.00000000
27	0.000	0.00000000
29	0.125	0.00000000
31	0.000	0.00000000
33	0.000	0.00000000
35	0.000	0.00000000
37	0.000	0.00000000
39	0.125	0.08333333
41	0.125	0.08333333
43	0.000	0.00000000

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 41

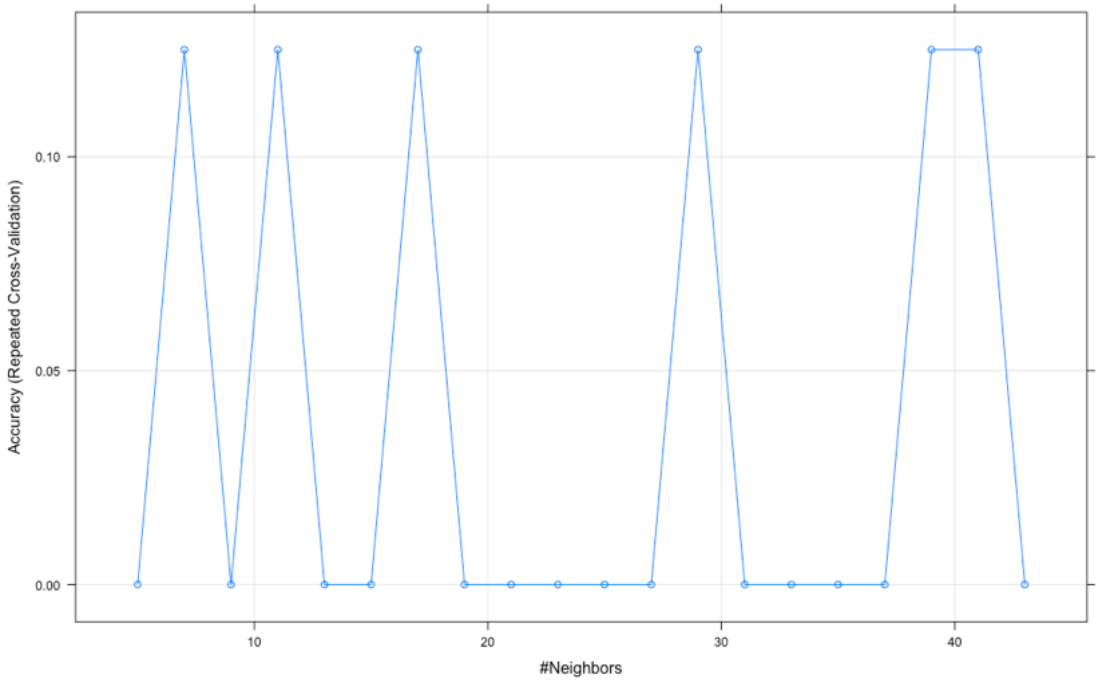


Figure 42. KNN Model Performance on the Cluster Analysis results taken from K-means: Accuracy

Likewise, in the above KNN cluster analysis console results and figure, it is verified that the model that was used was in fact k-Nearest Neighbours. Again, similar in comparison to the two previous datasets, the sample is 12 and the predictor is 2, the resampling: Cross-Validated (10-fold, repeated 3 times), likewise as stated above for each cross validation they were split 10 folds or parts, so nine of the are used for creating the model and the final one is used for accessing the model. It yields can see the: k Accuracy and Kappa results for the various values of k. Along with the summary of sample sizes and the model resampled results across tuning parameters. Mentioned again accuracy was used to select the optimal model using the largest value. The final value used for the model was k = 41. Therefore, while using KNN with the cluster analysis dataset results for K-means the best or optimal K or the k value to use for the clustering analysis dataset used was 41, the optimal model for KNN would be when k = 41 showing these results: 0.125 and 0.08333333. Therefore, again for

the clustering analysis dataset making the k value 41 the most recurrent label within the clustering analysis dataset that was performed in the classification KNN model.

In Figure 41 which showcases the KNN model for the cluster analysis from the K-means algorithm plotted in RStudio. As it is shown in the figure the x axes were labelled as the K Neighbours with the highest value point the ending is 41. Thus, the y axes were labelled as the accuracy, which again is the repeated cross validation, as a reminder each cross validation they were split 10 folds or parts, so nine of the are used for creating the model and the final one is used for accessing the model. Same to the previous Figure, 38 the k results are staggering throughout the line graph having high, highs and low, lows. Essentially, the features of the line graph in RStudio which plots the above KNN model (as well as the others shown above) makes the results clearer for the viewer.

6.7 Summary

In conclusion to summarise, this chapter describes the K-Nearest Neighbour algorithm that was used to analyse the results of the K-means algorithm to require and address the research questions and to test the hypothesised relationships developed in this study, cyberstalking indicative content on social media platforms such as Twitter. As well as, seeing KNN in comparison to K-Means and its results. The chapter begins with a discussion of the K-Nearest Neighbour Algorithm, while KNN is a supervised learning algorithm used for classification and while the model preforms the most optimal value for k is given. The data is assigned to the class which has the nearest neighbours. Followed using R Programming to assess the algorithm, the results that were used were the emotional terms results, the overall dataset results, and lastly the cluster analysis all from the K-means results within chapter 5. As seen, throughout this thesis there are many reoccurring themes within this study and

matching the two algorithms to see each result are imperative towards this study and future work within this field.

The chapter then continues with descriptions of how KNN was used on the data that was collected and its results from the clustering algorithm. Along, with how KNN is used within academia and why KNN is such a benefit for this research. Next, the reasoning as to the choice of algorithm and the analysis of the results are discussed. In addition, the limitations and recommendations for future studies are suggested and mentioned, as well as what are expected to achieve and potentially continue to do after this study is concluded. It is important to understand and realise that this type of study has not been done before within academia. Therefore, all the findings or results as well as, the datasets that are being used are so impactive to the cause and narration of cyberstalking on social media, mainly Twitter. Again, this research sets the tone for this type of work within academia, the narrative within academia towards cyberstalking on Twitter is very vague.

More importantly, this research will help open that conversation and shed light on cyberstalking indicative content on Twitter, dating mining with tweets in correlation to cyberstalking, with the use of KNN on the results of the K-means. All of the methods and results that have been concluded in this study, consequently prove, the findings in this chapter help corroborate one of the research questions being asked or in comparison: “Which data-mining algorithm is better suited for identifying and detecting cyberstalking on social media platforms?

Chapter 7: Discussion and Conclusion

7.1 Overview

This final chapter concludes the research aims and objectives that were outline in the introduction chapter. Along with the research questions themselves and a breakdown of what the findings and results are described in detail. As well as the discussions to each research question and which chapters highlight and correlate the correct information for each question. In addition to, how the research was conducted and why it was conducted that way. Moreover, how the results of each chapter help corroborate the research focus or aims as well as its importance. Furthermore, any limitations or recommendations to the current study and lastly, the future work in relation or continuous with this study or having new ideas for a new study overall, but with the same focus cyberstalking.

While discussing this research and its findings, it is important to refresh on the information used for the research topic. For instance, as mentioned in *Chapter 2: The Literature Review,* which gave a brief definition to cyberstalking: cyberstalking is the repeated unwanted relational pursuit of an individual through communication technologies, such as computers, tablets, and smart phones (Goodno, 2007; Reyns et al., 2012). Internet technologies are enticing platforms for stalkers because they create unique opportunities for perpetration (Nobles et al., 2014; Reyns, Henson, & Fisher, 2011). Although, cyberstalking can have numerous definitions. Gibson (2019), defined cyberstalking as follows: stalking via some form of electronic medium such as email (Finn, 2004; Fox, Nobles, & Fisher, 2016; Strawhun, Adams, & Huss, 2013), or social media platforms such as Facebook and Twitter (Bennett, Guran, Ramos, & Margolin, 2011; Fox et al., 2016; Henson, Reyns, & Fisher, 2013; Marcum Higgins, &

Ricketts, 2014; Nobles, Reyns, Fox, & Fisher, 2014; Reyns, Fisher, & Randa, 2018; Strawhun et al., 2013). In addition, as previously mentioned cyberstalking is the stalking of another through methods of electronic access and communication, such as, with the use of hidden webcams, GPS devices, and Spyware to monitor victim's behaviour, and pursuit and contact under anonymity through fake online pro-files (Sheridan & Grant, 2007; Shorey, Cornelius, & Strauss, 2015).

Machine learning is a major key within this study and is discussed throughout this thesis. Just mentioned machine learning is a huge important factor towards this study. Machine learning algorithms are programmes that can learn from data and improve from experience, without human intervention. Learning tasks may include learning the function that maps the input to the output, learning the hidden structure in unlabelled data; or ‘instance-based learning’, where a class label is produced for a new instance by comparing the new instance (row) to instances from the training data, which were stored in memory. “Instance-based learning’ does not create an abstraction from specific instances.

There are two learning forms for machine learning. They are supervised and unsupervised Learning, defined by Ted Coombs (2018):

- **Supervised learning**: supervised learning is exactly what it sounds like. Someone supervises the input of information upon which the learning algorithm will arrive at a conclusion. Think of this like giving the computer a tutor. One of the most basic supervised learning algorithms is designed around a decision tree. This is the foundation of the expert system, a series of yes and no questions sufficient for the computer to arrive at some probable answer. With an expert system, a conclusion is derived based on the programmed inputs of field experts. For example, diagnosing starter problems in a car will require the

user to answer questions about the symptoms experienced when trying to start the car. Do you hear a click when you turn the key? Yes or No. Based on that answer, new questions along the tree are asked until the computer suggests, “Your battery is likely dead.”

- **Unsupervised learning**: unsupervised learning allows for the training of AI, using data that’s unlabelled and unclassified with the use of special algorithms that allow the AI to learn on its own rather than being spoon fed the data by a human. Two common unsupervised algorithms include the apriori and the k-means (Coombs, 2018).

Within this study two algorithms were used to perform on the datasets for this study. The first algorithm that was used and seen in *Chapter 5: Clustering Algorithm.* K-means is an iterative algorithm that groups similar data into clusters. It calculates the centroids of k clusters and assigns a data point to that cluster having least distance between its centroid and the data point. Here’s how it works: start by choosing a value of k. for example, use k = 3. Then, randomly assign each data point to any of the 3 clusters. Compute cluster centroid for each of the clusters. K means clustering is an unsupervised learning approach to machine learning. Clustering is a useful tool in data science, it is a method for finding cluster structure in a data set that is characterised by the greatest similarity within the same cluster and the greatest dissimilarity between different clusters. Hierarchical clustering was the earliest clustering method used by biologists and social scientists, whereas cluster analysis became a branch of statistical multivariate analysis (Sinaga and Yang, 2020).

The second algorithm that was used within this study introduced in *Chapter 6: K Nearest Neighbour.* KNN also known as: K-Nearest Neighbour Algorithm. KNN

classifier is one of the most common and easy to implement classifier in the machine learning domain, achieving competitive results compared with most complex methods, and sometimes it is the only available choice, for example when used for content-based image retrieval Hassanat (2018).

However, Hassanat (2018) continues, with mentioning how KNN is a very slow classifier and a lazy learner. For instance, testing any example, the KNN classifier cannot produce a small fixed-size training set of *n* examples, in *d* dimensional feature-space, the running cost to classify one example is O(*n.d*) time, Hassanat (2018) submitted: for the blessing or curse of big data, where n and/or d are relatively large values, big data sets includes their ability to provide a rich source of information to the classifiers for a better learning, while the curse of big data sets includes their very large sizes.

7.2 Introduction

The aim, purpose, and or objectives of this study was to showcase all the previous studies of social media analytics (data mining-based) that are reported in the open literature focus on cyber fraud, cyber bullies, and cyber hate crime. As well as bring forth cyberstalking analytics which has not been given great attention by the researchers in the past and thus which motivated this study. In addition, lightweight data mining algorithms have not been used to detect cyberstalking on social media platforms with the use of Twitter (Karyofyllis, 2018). Moreover, the aim of this PhD research is with the use of data mining and machine learning, to have security metrics to detect cyberstalking from social media platforms with the use of Twitter.

In addition, as mentioned above and in the introduction, the detection of cyberstalking, harassment, and security threats on Twitter was conducted by using the data mining analytics along with the algorithms and machine learning being used.

The derived security metrics are then used to flag up any suspicious cyberstalking content (text-based and/or audio-based), to detect and prevent potential cyber stalking on social media platforms. With the expansion of the Internet, harassment, abuse, and threats increase in volume, velocity, and language. As such data mining and analytics can help detect the rise in harassment and threats that are part of cyberstalking.

Lastly as mentioned, there is indeed a gap within the literature within this field. Mentioned throughout this thesis, similarities to cyber bullying or cyber harassment are in comparison to cyberstalking. However, cyberstalking is not given the ample light or the noticeable recognising that is needed on this extensive topic. The conversation around cyberstalking with the use of this social media platform that was used for this thesis, is not happening within academia, unless cyber bullying or cyber harassment is attached to the topic at hand. Now, with that is not being said that there is a narrative for this topic, but not a stand-alone narrative, in connection with the social media platform, Twitter. Moreover, as mentioned within the literature review, the knowledge gap that surrounds cyberstalking is assisted with previous research helps to develop a greater understanding of cyberstalking such as what legally constitutes cyberstalking, the role society plays in governing cyber-based misbehaviours, and regulatory issues governments and institutions face when attempting to prevent it. However, the literature does not comment on how actual measures can be developed to detect cyberstalking.

Previously stated above, was how and why this research was designed and its indented purpose. Moreover, how the research was conducted and structured. This thesis had the aim of exploring these research questions at hand and the four questions that the researcher is using for this study are as follows:

1. How can data mining and quantitative analysis of random open-sourced data samples reveal cyberstalking indicative content on social media platforms?

2. What security metrics indicate whether cyberstalking has been developed through social media platforms?

3. How can these metrics be used to provide a fine-grained measurement of cyberstalking?

4. Which data-mining algorithm is better suited for identifying and detecting cyberstalking on social media platforms?

7.2 Research Findings and Innovations

7.2.1 First Research Question

"How can data mining and quantitative analysis of a random open-source data sample reveal cyberstalking indicative content on social media platforms?"

This research and its correlating study are formed around the aims and objectives, but more importantly around the four research questions at hand. *Chapter 3: Preliminary Data: Automatic Identification of Cyberstalking on Twitter using NVivo.* Coding is the prime focus for the first research question*: "how can data mining and quantitative analysis of a random open-source data sample reveal cyberstalking indicative content on social media platforms?"* In addition, *Chapter 4: Twitter Data Analysis with the use of R Programming/ R Studio.* Can also be used to answer the first research question as well. However, first looking at Chapter 3, NVivo with the google extension NCapture, both were used for collecting a sample of tweets from Twitter that had connections to cyberstalking. However, before the collection was obtained, the researcher viewed tweets in relation to cyberstalking to understand the

narration around the topic. Hence, at the start in the beginning three Twitter threads were searched on Twitter: “cyberstalking, #cyberstalking, and stalking and fear”. Now, the reasoning as to why cyberstalking and #cyberstalking was used was to see if there was any difference in the description of cyberstalking on Twitter within those two threads. Also stalking and fear were investigated based on the interpterion surrounding those terms in correlation to cyberstalking. Moreover, the importance of the data mining method that was being used to measure parameters such as:

- Terms / key words
- Number of postings / conversation or connections
- Probabilities (key words appear)
- Weightings of terms or key words
- Location of postings or connections (IP address)

Keeping these five details in mind, motivative how these three Twitter threads were used to locate the terms\ key words used with in the narration or conversation with regards to cyberstalking. The number of posting or connections towards cyberstalking overall is very imperative, because it lets the researcher known that indeed cyberstalking indicative content is happening on Twitter and how the numbers within key terms or unigrams are fluctuating. The weighting of terms or key words again is beneficial for this study to identify which terms or key words are more important to focus on while conducting this research. The location and postings or connections (IP address), is standard for this type of data mining machine learning software. However, all the information collected within that area was not used or needed for this study, keeping the study and its research anonymous. Importantly, these data mining parameters moulded how the research and study was going to be

conducted and analysed. Therefore, the next part of the study and the findings are very critical.

After, each of the three Twitter threads that are mentioned above was collected within NVivo using Ncapture, a word frequency and a text search were done on each Twitter thread. Thus, the core finding of the fifteen unigrams or key terms that are used as the prime innovation in this study within correlation with cyberstalking. The unigrams below are the framework of this study, without these fifteen unigrams the study and its results would be less informative towards cyberstalking indicative content on social media platforms and in this case Twitter. Below are the fifteen unigrams that were found and used within this study:

- Abuse
- Annoying
- Creep or Creepy
- Fear
- Follower or Follows
- Gender
- Harassment
- Messaging
- Relationships P/P
- Scared
- Stalker
- Technology
- Unwanted
- Victim

- Violent

While obtaining these unigrams the researcher took into consideration the narrative around cyberstalking on Twitter. Meaning, that he or she noticed certain familiar terms and correlations that would be linked to cyberstalking, he or she did not used the more common familiar terms in connection to cyberstalking within this study. For instance, some of those terms are paedophile or paedophilia, aggravated, trouble, hacking, and slander to name a few. Thus, the fifteen unigrams or key terms prove that cyberstalking indicative content in fact is on the social media platform: Twitter. These unigrams have multiple tweets that have a relationship with cyberstalking. Each unigram was searched in Twitter, with no advanced search just on its one (with no hashtag). With the use of the open-sourced social media platform and the programme NVivo, to conduct this part of the research correlated with the first research question*: "how can data mining and quantitative open source of a random open-source data sample reveal cyberstalking indicative content on social media platforms?".*

Each of the unigrams had 5000 tweets that were sifted through and the tweets that showed or proved to have cyberstalking indicative content were coded and used for the study. For instance, showing below the unigrams and the number of tweets that corroborate cyberstalking are the findings for the unigrams and tweets 5000, therefore proving: in fact, how data mining and quantitative analysis of a random sample open-source data sample reveal cyberstalking indicative content on social media platforms. Along with results of the preliminary dataset: the fifteen unigrams and the tweets correlate to prove the question at hand.

Unigram	Tweets (5000)
Abuse	95
Annoying	29
Creep or Creepy	150
Fear	60

Follower or Follows	41
Gender	41
Harassment	46
Messaging	41
Relationships P/P	44
Scared	35
Stalker	139
Technology	43
Unwanted	50
Victim	40
Violent	21

Preliminary Data Unigram/ Tweets (5000) Dataset

However, while using the software programme NVivo helped catapult this start this research and the vast results that were obtained during this process let alone as this work and thesis revolve. Whilst using NVivo was beneficial and less time consuming than other methods which helped the research move along efficiently gathering vast amounts of data that was informative towards the research topic and its questions. Having used NVivo brought an insight and innovation into the world of cyberstalking on Twitter, as well as associated the foundation of this research. Without the use of NVivo and NCapture, this study would have not had the insight that it needed to get the research process started and to obtain the valuable preliminary dataset that is used on many occasions throughout this study. Therefore, using data mining and machine learning programmes helped pave the foundation for this research and assisted with the validation of the rest of the research and its findings and innovations. Furthermore, the reasoning for the intention to the use of data mining techniques on social media is that the data is the empowering factor for advanced search in search engines such as Twitter and helps in better understanding of the data for this research. Lastly, it is imperative to know that each chapter elaborates more and more on this research question. Continuing to prove that cyberstalking inductive content is happening for instance, Chapters 4, 5, and 6 along with their findings and

innovations each of those chapters touch on this topic and bring more insight onto the first research question. Thus, uphold more so that in fact cyberstalking indicative content does in fact happen on social media platforms, such as Twitter in this case.

7.2.2 Second Research Question

"What security metrics indicate whether cyberstalking has been developed through social media platforms?"

Moreover, considering the second research question *Chapter 3: Preliminary Data: Automatic Indemnification of Cyberstalking on Twitter using NVivo Coding* and *Chapter 4: Twitter Data Analysis with the use of R Programming,* both answer the first research question, but the second one as well: "what security metrics indicate whether cyberstalking has been developed through social media platforms?". Considering the main goal of any security metrics is to assess how well an organisation, a network, or personal profile is reducing security risk. There are also different metrics that can provide insight into the performance of the program that is being used itself.

Therefore, in Chapter 4, the introduction to the second dataset that would be used is the random sample data set was brought into the study. This dataset was made up of five csv files that contain 1,500,000 data points in total. The security metrics that indicate whether cyberstalking has been developed through social media platforms, in this case Twitter. Would be the unigrams that were presented earlier. As well as tweets that correlate towards cyberstalking found within the random sample data set. Using the unigrams as the main source in connection to cyberstalking indicative content seeing if they have any relationship to the random sample dataset is vital. However, the methods that were introduced in Chapter 4 bring forth more insightful findings towards the research question that is in question.

Within Chapter 4 the method that was used was R Programming. For this research, the perceptive for the use of R Programming, which is beneficial because,

R programming is being used within academia to showcase further findings and research for the academic community. R Programming is a remarkable tool that helps correlate the task at hand, the security metrics that indicate whether cyberstalking has been developed through social media platforms and in this case Twitter.

For instance, an advanced search within R Programming was done to continue the analysis on the dataset. As can be seen in the appendix section from Dataset 3.csv (11 Tweets) which was taken from the appendix section to show the impact of advanced search has on this study. The focus is on the four out of fifteen unigrams that had tweets indicate or have a connection to cyberstalking indicative content. Those unigrams are abuse, creep/creeper, follow/follows, and lastly stalker. Therefore, abuse had one tweet with regards to cyberstalking. The unigram creep/creepy has four tweets viewing cyberstalking and follow/follows has one tweet regarding cyberstalking. Lastly, the final unigram that is seen is stalker which has five tweets concerning cyberstalking. Consequently, using R programming with the help of an advanced search correlated with the first research question as well as this one, the second research question. Proving that the unigrams are in fact informing the researcher that cyberstalking indicative content is seen thoroughly on Twitter. Thus, corroborating that the security metrics which can help detect, deter, and even prevent cyberstalking on social media with either the flagging or usage of the unigrams in comparison with other key terms, or with the quantity of times each unigram is used within the social media platform. All the vast outcomes can tie into the unigrams themselves and how they are showing a relationship with cyberstalking.

In addition, it is essential to notice that the unigrams that were picked from the preliminary data are in fact are useful and can potentially detect cyberstalking tendences on Twitter. Linking all the results and findings together which make great

innovations for this research. Which all the innovations in each chapter are vital for this study. Moreover, it is remarkable looking over each dataset within the appendix that some of the same unigrams are flagged more than once with the tweets associated towards cyberstalking. Furthermore, within the advanced search looking at the potential metrics of which unigrams are being flagged up and how many times each are, and which ones have not been brought forward.

Starting with dataset 3 csv the unigrams that are flagged with tweets in relation to cyberstalking are: abuse, creep/creeper, and stalker. For dataset 1 csv the reoccurring flagged unigrams are: annoying, creep/creeper, follow/follows, and stalker. In dataset 2 csv the unigrams are: follow/follows, stalker, and technology. Dataset 4 csv repeated unigrams are: annoying, follow/follows, and stalker. Lastly, in dataset 5 csv the repetitive unigrams are: abuse, creep/creeper, and stalker. Now, looking at these findings the constant is the unigram stalker which has been viewed in each dataset and the unigram that has the most tweets in relation to cyberstalking, with a total of 20 tweets between each dataset. However, that isn't shocking since the unigram stalker is the closest related term to cyberstalking. On the other hand, the remaining unigrams: abuse, annoying, creep/creeper, follow/follows, and technology are repeating but not as high volumes as stalker. For instance, abuse is seen in two of the datasets with a total of 3 tweets between each dataset. Annoying is seen in two datasets as well but has a total of 2 tweets. As for creep/creeper which is present in three datasets and has a total of 9 tweets. Lastly, the unigram technology was the only unigram present in one dataset and has a total of 2 tweets.

Furthermore, with a brief introduced in Chapter 4 the use of emotional terms taken from the random sample dataset. These terms were looked at and analysed in *Chapter 5: Clustering Algorithm*. The emotional terms that were used are "bad", "sad",

and “hate”. Importantly, these terms were first seen and noticed as stated in Chapter 4. While using R Programming to obtain results within the random sample dataset. Whilst plotting those results which was the last step within the process these terms were seen in nearly every csv file dataset that was run through R. Two out of the five datasets did not have the emotional terms “hate” which are dataset 4 and 5. As seen in Chapter 4, Figure 29, “sad”, “bad”, and “hate” are seen in the figure. Now, close analysis of this figure we can see that it is taken from dataset 3 csv file: “sad” as the highest count of 1470 and the highest rank of the three at 23. Followed, by “bad” in second with a count of 1338 and a rank of 29 and lastly, “hate” with the lowest of the three of a count 1048 and rank of 41. As well as dataset 2 seen in Figure 30, in Chapter 4, “sad” having the highest count of 1695 and rank of 19, “bad” followed with a count of 1312 and a rank of 34, and lastly “hate” with a count of 1104 and the rank of 40. In the appendix section there are all the remaining tables and figures. However, viewing each dataset and its results are fascinating for instance, dataset 1, which has the same results as dataset 2: “sad” has a count of 1695 and a rank of 19, followed by “bad” with a count of 1312 and a rank with the rank of 34; lastly, “hate” with the count 1104 and rank of 40. In dataset 4, reminder “hate” was not in the results: “sad” has a count of 1856 and the rank 14, along with “bad” having a count of 1313 and a rank of 33. Lastly, dataset 5 (same as dataset 4 “hate” was not in the results): “sad” having a count of 1725 and a rank of 17 and “bad” with a count of 1329 and a rank of 31. Lastly, the ranking of these emotional terms are beneficial for the 4th research question and with K means algorithm used furthers the research.

Again, since the main reason as to why security metrics is used is to prevent or reduce a security risk. With the results from the preliminary dataset and the random sample dataset prove that cyberstalking indicative content does in fact happen on

Twitter. However, with the above findings prove that cyberstalking content is widely being active on Twitter. With the use of the unigrams and the advanced search in R proves that Twitter is imperative to security metrics. Therefore, the awareness of the intended identification of a potential risk which would be cyberstalking whether to an individual or a group on social media platforms, can possibly be identified within this in this study with the use of Twitter. Moreover, this question does give a brief insight to the third and fourth research questions.

7.2.3 Third Research Question

"How can these metrics be used to provide a fine-grained measurement of cyberstalking?"

Whilst answering the previous research question gave insight on how and what the security metrics are for this study in correlation to which chapters helped answer the question. Thus, the metrics can be used to provide a fine-grained measurement of cyberstalking mainly on how the metrics help determine in detail the awareness of the cyberstalking indicative content as well as how to possibly prevent it from happening in the future. For instance, the breakdown of each advanced search which was done in Chapter 5 along with the unigrams that are reparative and shown in many tweets can help with the detection process. Meaning having the focus on those terms as well as the emotional terms mentioned in Chapter 5, that were used within the analysation process throughout this study. Moreover, the importance of ranking, the ranking of unigrams which unigrams have more tweets and narration around them compared to others. The ranking of the emotional terms individually was seen through the analysis process. More so, the ranking of each the unigrams and the emotional terms together. Along with the tweets that each recorded separately as well as together. Alongside with the original reoccurring themes from the preliminary dataset that are focused on in Chapter 3 can help correlate how these metrics can be used to help deter

cyberstalking suggestive content on social media platforms. As well as with the use of the algorithms that are mentioned in detail in Chapters 5 and 6, which were used in this study help pave a new conversation on how different angles and aspects of cyberstalking indicative content can be not only shown but analysed to help further this on-going issue on social media platforms.

Moreover, this question is a continuation of the last research question taking into consideration the findings that were mentioned for the security metrics themselves, thus the answer of how those security metrics can make a fine-grained measurement of cyberstalking. For example, looking at the ranks of emotional terms in more detail if “sad” is in all the five random sample datasets csv files and has the highest count and rank over the remaining two emotional terms then possibly that the emotional term can be investigated more frequently in correlation to detect cyberstalking. However, that does not mean the other two emotional terms are not as important, they are in fact, looking at the emotional terms paired with the unigrams and the tweets it is fascinating which emotional term along with the unigram have the highest volume of tweets. Such as, seen in Chapter 5, there were fifteen tweets taken from Chapter 3, from the preliminary dataset results which were used to illustrate the connection to unigram and emotional terms. As can be seen in those tweets, the emotional term “sad” is in 4 out of the fifteen tweets with the unigrams: technology, creep, follow (following), and relationship. Finally, it yields, the unigrams vs. the tweets 5000 the unigram creep was the top unigram that had the most tweets. Therefore, a possible security metric that can be developed would be using the unigram creep and the emotional term “sad” together to see if there any indication of cyberstalking activity on Twitter or other social media platforms. However, as previously mentioned the emotional term “sad” being the top term out of the others. It is not the top emotional

term that is seen in these fifteen tweets. Likewise, both emotional terms “bad” and “hate” are shown respectively “bad” in six out of the fifteen tweets and “hate” 5 out of the fifteen tweets. Along with the unigrams: annoying, creep, fear, messages, stalker, unwanted, technology, violent. Consequently, again, in Chapter 5, the clustering results were used within tweets as well.

Similarly, to above there are 10 tweets in total from the clustering results again within Chapter 5, that are mentioned each tweet has again a unigram and an emotional term attacked to it. Such as, above all three emotional terms are used “hate” being seen in 4 out of the 10 tweets. Then “sad” being seen as well 4 out of the 10 tweets and lastly, “bad” being seen three out of the 10 tweets provided from the cluster results. Now, some of the emotional terms like “sad” and “hate” are seen in the same tweet. However, the unigrams that are attacked to each term are for “hate”: stalker, creepy, technology, and scary (scared). For “sad”: abuse, scary (scared), creep, and stalker. Finally, we have “bad”: fearing (fear), creepy, and stalking (stalker). In addition, knowing that the emotional terms “bad” and “sad” are shown within 100 percent of the clustering results and 100 percent of the word frequencies results of the random sample dataset csv file. While “hate” is seen within 60 percent of the clustering results and 60 percent of the word frequencies results of the random sample dataset csv files.

Comparing these results together it correlates that the security metrics of ranking each emotional term as well as the unigrams. Help prove or correlate that cyberstalking suggestive content is in fact on Twitter, but also helps push the focus on to how those security metrics can perform a fine-grained measurement of cyberstalking. The unigrams that the researcher suggested and used within the preliminary data set in fact corroborate that detection of cyberstalking on Twitter is admissible with their use. Furthermore, the emotional terms: “bad”, “hate”, and “sad”

beside the unigrams are attached to the tweets mentioned within this study from both data sets, proving that these emotional terms and unigrams go hand in hand with regards to cyberstalking indicative content on Twitter. Together with, with the results within this thesis along with all the methods that have been used and with structured datasets verifies that all the findings within this research linked together showcase that this research has sustained true to its objectives and aims.

7.2.4 Fourth Research Question

"Which data-mining algorithm is better suited for identifying and detecting cyberstalking on social media platforms?"

As mentioned in *Chapter 6: K Nearest Neighbour*, in the present-day scenario, machine learning and artificial intelligence are replacing all the conventional computational techniques and programming languages, most importantly machine learning gives computers the ability to learn without being explicitly programmed. Therefore, that is why two known machine learning algorithms were developed in this study to help aid the research at hand and to further the results that were obtained.

7.2.4a K Means

In this thesis there were two data-mining algorithms were developed within this study on the datasets provided. As mentioned in the previously two chapters within this thesis, Chapter 5, and Chapter 6. The two algorithms were K Means and KNN. As in seen *Chapter 5: Clustering Algorithm*, the introduction of the first algorithm is brought forth the algorithm that is used first is K-Means. K means clustering algorithm is generally the most known and used clustering method. As mentioned within the chapter: K means clustering is extensively used in various fields such as text mining, machine learning, image analysis, image processing, web cluster engines, bioinformatics, weather report, and so on. Therefore, that is why k means was used

for this study. It has been shown that k means is used in various fields and two of those fields are the main methods of this research: text mining and machine learning. Again, as stated in Chapter 5, the main goal of the current research is to visualise the linking of the results from the k means cluster algorithm along with the unigrams; to showcase the correlation or relationship towards cyberstalking indicative content within tweets on the social media platform Twitter. As well as it envisions the clustered tweets or terms according to the relationship of cyberstalking.

Moreover, it is important to remember all the key findings thus far and the linking of the reoccurring themes within the findings. Such as, the *unigrams that were found and presented in the preliminary dataset. Likewise, the unigrams were also presented in the random sample dataset tweets.* Also, those unigrams were then found in pre-existing tweets within the random sample dataset and have what was introduced as the three emotional terms attached to them. The advantages of clustering programming as compared to other sources is insightful. Not to mention with the use of R programming which gives k means algorithm the strength to compete with its competitors. For instance, what Chapter 5, brings forward is how k means clustered that data into groups, which the data itself was not pre-grouped. Then k means determines how many groups or clusters are within the dataset. Again, this entire process was done to each random sample data set (5 csv files) within RStudio. Furthermore, the research did not use k means for his or her work, he or she would not have the same profound results or would not have the same advances with cyberstalking. K means was the biggest part of the analysation process within this research. The findings alone speak for themselves for instance without k means all the input that is gathered from the algorithm would not be present which was beneficial for this research.

As mentioned in many of the chapters in this thesis, emotional ties to cyberstalking have been seen within the preliminary dataset and the random sample dataset. As well as how as found in *Chapter 4*: “that there is a similarity to emotional terms that can be used within or as an exception to cyberstalking”. Therefore, the emotional ties are one of the focuses alongside the unigrams whilst using k means. While answering another research question, the emotional terms, or links to the ties towards cyberstalking was used for the research question as well. Within this study or research there are many parallels and the same goes for the research questions. However, as the raw random sample dataset is the prime dataset for k means. As shown in *Chapter 5*, data set 3.csv file: K-means clustering with 12 clusters of sizes 1688, 3155, 1229, 2176, 1604, 2092, 2381, 3259, 27883, 2073, 1534, 926. In addition, with these clusters the emotional terms: “bad”, “hate”, and “sad” all are within these clusters. “Bad” is in the first cluster, followed by hate which is in the fifth cluster, and lastly sad which is in the ninth cluster. It is interesting that each dataset had different results again from Chapter 5, data set 1 all three emotional terms were in one of the twelve clusters: “bad” again was in the first cluster, “hate” was in the fourth, and lastly “sad” was in the sixth. Data set 2: “bad” again was in the first cluster, “hate” was in the eighth, and lastly, “sad” was in the fifth. However, for data sets 4 and 5 only two of the terms were in one of the twelve clusters. The two terms: “sad” and “bad” were seen in one of the twelve clusters; for data set 4 “bad” was in the first and sad was in the seventh and for data set 5 “bad” again was in the first and “sad” was in the sixth. Moreover, the emotional term “bad” was in the first cluster for all the datasets which seems appropriate for this research.

Additionally, while using k means broke down the ranking of each emotional terms from the cluster analysis preformed on each dataset csv file from the random

sample dataset. As well as k means gave the opportunity to have cluster dendrogram for each dataset csv file as seen in *Chapter 5*, in Figure 32, Cluster Dendrogram: Data Set 3. This figure correlates the argument that the emotional terms are in fact showcased in the random sample dataset hat a high volume. Within the figure it illustrates “sad’ has the highest rank in comparison with the other two terms and “bad” and “hate” are closer together on the contrast. Therefore, proving these terms are constantly showing persistence throughout each data collection and the analysation process. Besides, proving that the emotional terms are a constant force within the random sample dataset. Also, shown in tables 2-6 in *Chapter 5*, are the cluster results for the data sets for the random sample data set (all five) are included. For example, dataset 1: “hate” was ranked the highest, followed by “sad” and then “bad”. Dataset 2: “sad” was ranked the highest, followed by “bad”, and lastly “hate”. Dataset 3: “sad” was ranked the highest, followed by “bad”, and then “hate” being third. Remember Dataset 4 only has two out of the three emotional terms within its cluster results: “bad” is the higher of the two followed by “sad” and finally dataset 5: “bad” is ranked the highest and “sad” is second. Moreover, while using k means broke down the ranking of each emotional terms from the cluster analysis preformed on each dataset csv file from the random sample dataset.

Another benefit that k means brings forth the three emotional terms are shown with the sum of the total of each cluster within each data set, which is highlighted in yellow. As well as the average of the total the sums of each cluster from each data set, which is highlighted in orange. As previously stated, the emotional term “bad” was constantly in the first cluster within all five data sets. Therefore, making these results surprising to be the second strongest sum or value of the three terms with: 3.175042124 followed by “hate” with: 1.401795714 and lastly the strongest sum or

value emotional term: “sad” with: 4.155972524. These findings with the use of k means are so insightful for this research with the cluster ranking of the emotional terms and proving they are a focus within the random sample dataset as well as the preliminary dataset.

K means was not only used for the emotional terms, but it was also used for clustering results within tweets. The prominence or understanding of how these three emotional terms and the fifteen unigrams showcase that there is cyberstalking indicative content on Twitter is informative for this research. As the findings and results are brought together it is noticeable that the unigrams and these emotional terms go hand and hand with each other. In fact, both findings are imperative towards the research at hand, however, they both are very impactful with regards to cyberstalking suggestive content either together or apart but more so together.

This research needed to be imperative to the fact that cyberstalking indicative content is in fact on Twitter. Also, how that content can be detected and how it can possibly even be prevented. Thus, using k means clustering to strengthen the research and gather all findings that are beneficial for the research to address the research questions.

Again, showing that there are many parallels with each chapter and the research questions. For instance, while answering research question 4: the 10 tweets from Chapter 5 were used to help corroborate that research question. Also, it could be possible that without using k means a few of the other research questions could not be answered. For instance, k means ranked the emotional terms also linked the emotional terms and the unigrams. Then verified that both together the unigrams and emotional terms are within both datasets: the preliminary dataset and the random sample dataset. Therefore, showing the outstanding findings brought forward by k

means. These clustering results within tweets substantiate that the unigrams, emotional terms, and both the preliminary datasets and the random sample datasets (all five csv files) go hand in hand with each other and cyberstalking. Precisely, making k means the best algorithm to use for this study; and answering: "*which data-mining algorithm is better suited for identifying and detecting cyberstalking on social media platforms?"* However, KNN was to further the findings brought forward by k means clustering algorithm and the findings from KNN brought onwards many insights and findings.

7.2.4b K Nearest Neighbour

As stated before, KNN (K Nearest Neighbour) was used for this study in comparison to k means and its results for the emotional terms. KNN algorithm first finds k nearest neighbours of a query in the training dataset, and then predicts the query with the major class in the k nearest neighbours. Furthermore, since K-nearest neighbour method is a well-known classification algorithm used in pattern recognition. It was used on the results taken from chapter 5, the emotional terms results. Moreover, it is important to know how KNN differs from K means, a supervised machine learning algorithm like KNN (as opposed to an unsupervised machine learning algorithm like K-means) is one that relies on labelled input data to learn a function that produces an appropriate output when given new unlabelled data. Likewise, an unsupervised machine learning algorithm makes use of input data without any labels in other words, no teacher (label) telling the child (computer) when it is right or when it has made a mistake so that it can self-correct. Unlike supervised learning that tries to learn a function that will allows people to make predictions given some new unlabelled data, unsupervised learning tries to learn the basic structure of the data to give people more insight into the data. Lastly, the significances of the types of

classifications that are used within KNN. There are two types of KNN algorithms that can be used, one is classification, and the other is regression. KNN classification algorithm first selects k closest samples (i.e., k nearest neighbours) for a test sample from all the training samples, and then predicts the test sample with a simple classifier, e.g., majority classification rule. As for regression the KNN regression has been widely used and studied for many years in pattern recognition and data mining. Moreover, the KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood.

As seen in *Chapter 6 K Nearest Neighbour*, the importing of the k means results was the first stage of the process for this chapter. Begging with importing the result from C*hapter 5*, was essential for KNN to work. After, the results were imported the classification steps were next. The reason as to why the classifications steps were performed within the RStudio console that the researcher used for the preparation for KNN. The first step was done for the purpose of imported the dataset that was used for the KNN algorithm. The second was to see the structure of the dataset in R and to see if anything needed to be changed. For example, the variable emotional.term needed to be changed due to the listing or ranking of each term. The variable emotional.term was the three emotional terms used: “bad:, “sad”, and “hate” they were originally stored in RStudio as a character vector and they needed to be changed to a numeric or factor vector for the KNN model to perform. Therefore taking the variable emotional.term for each emotional and making them correlate with the numbers 1,2,3 meaning bad=1, sad=2, and finally hate=3. Thus, labelling those variables as numeric/factor: 1,2, and 3 and not having them as character variables or having the overall tally of each variable separately. Lastly, the variable emotional.term was made

to a factor the reasoning as to why this was done was, for the soul resolution of the research type and dataset that were being used which were better suited for performance of the KNN model.

Furthermore, the final step before the model is prepared and run within RStudio. The data partition for this model the data partition had the set.seed to (1234) though the independent sample is used with size 2 and the number of rows which is (data). This is a sampling with replacement so replace=T (true) with a probability of 70 per cent for training data and 30 per cent for test data. Now, with the training the rows in the data with regards to independent sample to: == [1,]. The training data: which consists of 12 observations and 3 variables and the test data: which contains the following 1 observation and again 3 variables. The way the training and test data are set up this way is because, with the fact that the training data was 70 per cent and the test data was the other 30 per cent. As stated in *Chapter 6*, the researcher did try and have the percentage to 60 and 40, however the results were inconclusive therefore the reasoning for having the test data the training data set to 70/30.

All the above steps or directions were put into place for the KNN model to run its performance or analysation. The KNN model was the final process for the KNN algorithm to run. Moreover, to have the KNN model run properly the method or correct package to use is “repeatedcv” which is mentioned in *Chapter 6*, which is also known as the repeated cross validation which is mainly used for again mentioned in *Chapter 6*: create multi fold, the code iterates over multiple times (given by repeats in train Control () syntax in R) for each k cross fold (given number). In cross fold, while using CV, it is a one-time process on each of the fold (set by using numbers in train control (). Then the number 10 is used for recent iterations and the cross validation is set to repeat 3 times; the number of complete set of folds to repeat this validation. The

“repeatedcv” is the prime focus for the KNN model a few other details are submitted within strings for the model to run efficiently. Likewise, the choosing of the number and how many times each cross-validation run is important. For this dataset the number was set to 10 and the repeats was set to 3. The set.seed was set to (222), then the training dataset with the method being KNN, also adding a truneLength of 20, adding TrControl that was used and created earlier in the process, finally followed by the plot of the each dataset.

The model performance and its results are profound for this research and its innovations. As stated before, the KNN model was used the results from *Chapter 5*, which was on the overall dataset, the emotional terms, and lastly the cluster analysis. Therefore, each breakdown of the dataset was run through the KNN model separately. For instance, the results for the overall dataset are the sample is 12 and the predictor is 2, the resampling: Cross-Validated (10-fold, repeated 3 times), for each cross validation they were split into 10 folds or parts, so nine of the are used for creating the model and the final one is used for accessing the model. As well as it yielded the: k RMSE and MAE results for each potential k. The final value used for the model was k = 43. Therefore, while using the KNN model and with its performance and the overall dataset results for K-means the best or optimal K to use for this model was 43, thus, the optimal model for k value for the overall dataset used for KNN would be when k = 43 showing these results: 1.199764 and 1.189697.

The results for the emotional terms dataset as imperative since these emotional terms are a major asset to the research and were foreshadowed throughout the entire study. Slightly like the overall dataset, the sample is 12 and the predictor is 2, with 3 classes: '1', '2', '3'. These classes are the emotional terms (“bad” = 1, “sad” = 2, and “hate” = 3) again as the same as before, the resampling: Cross-Validated (10-fold,

repeated 3 times). As well as it yielded the: k Accuracy and Kappa results for the various values of k. Along with the summary of sample sizes and the model resampled results across tuning parameters. Accuracy was used to select the optimal model using the largest value. The final value used for the model was k = 11. Therefore, while using KNN with the emotional terms results from k mean the best or optimal K to use for this model was 11, the optimal model for KNN for the emotional term dataset that was used would be when k = 11 showing these results: 0.39855072 and 0.03703704. Lastly, the cluster analysis themselves were the last part of the results from k means used for the KNN model. Again, similar in comparison to the two previous datasets, the sample is 12 and the predictor is 2, the resampling: Cross-Validated (10 fold, repeated 3 times), As well as it yielded: k Accuracy and Kappa results for the various values of k. Along with the summary of sample sizes and the model resampled results across tuning parameters. As well as before, accuracy was used to select the optimal model using the largest value. The final value used for the model was k = 41. Therefore, while using KNN with the cluster analysis dataset results for K-means the best or optimal K or the k value to use for the clustering analysis dataset used was 41, the optimal model for KNN would be when k = 41 showing these results: 0.125 and 0.08333333. The importance as to why the KNN model is beneficial with its results from the k means algorithm is essential.

As stated in *Chapter 6*, KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression). The KNN model that was made and performed on the datasets that was used was a classification model. Looking at each dataset: the overall dataset, the emotional terms dataset, and lastly the cluster

analysis dataset finding the optimal k value is astonishing to this research. For example, the overall dataset the optimal k value within the KNN classification model was 43, therefore the best outcome or result for that dataset is when k reaches its highest prospective which is the value of k=43, making the value 43 the most recurrent label within that dataset. Likewise, for the emotional terms dataset the optimal k value was k=11, making the k value 11 the best optimal outcome for the most frequent label. Lastly, the clustering analysis dataset having the optimal k value be k=41, therefore the most frequent label within the clustering analysis dataset is having the k value = 41. While using KNN showcases the most frequent label within each dataset used from the k means results, it helps illustrate that there are reoccurring themes within this study. Since the KNN model used was a classification model which its main benefit is finding the values of the most frequent or reoccurring k values in each of the datasets provided.

The advantages of using machine learning algorithms are the efficiency, accuracy, performance, and usability to different domains of analysation. Both these algorithms bring forth many insights to the study at hand and for topic cyberstalking. Each algorithm is widely beneficial for their purpose and that was the reason they were chosen. However, as for the question at hand the fourth research questions with algorithm are better suited for the identification or detecting and preventing cyberstalking on social media platforms is in fact with this study and research K Means Clustering. The results that were obtained with k means were endless and very beneficial for the identifying cyberstalking and for a possible way to prevent it using the probable security metrics that can be found within the results. As well as the linking between the results using k means and the preliminary dataset that was used in the beginning of this research. The clustering algorithm was the best option for this type

of study and the paramount decision for the analysation of the datasets that were used within this study.

Nonetheless, with the use of KNN analysing the results from the k means clustering algorithm also aided the research. While using KNN as a second data mining or machine learning tool helps bring forward the profound results within this study. While looking at the clusters broken down into three different datasets, from the previous algorithm results at much more detail. More importantly, looking at both these algorithms one amplified and / or elevated the results of the other. Therefore, making them go hand in hand with this research and study. For instance, the reasoning as to why the KNN model was a classification model was to prove that there are frequent or reoccurring values within each of the datasets that were used from the k means results. Conversely, without the use of k means none of this could have been implied therefore, making k means the more imperative algorithm for this study. Additionally, using both the algorithms K means and KNN has opened the narration to a topic that is not widely seen in academia, however parts are seen like: cyberbullying, harassment, and even studies on social media.

7.3 Research Limitations and Recommendations

7.3.1 Research Limitations

As this study progressed some limitations on this form of research or study are inflicted within this study would be in within the design of the study and time efficacy would be a foremost limitation on this study. Mainly, because of how time consuming it is to gather material on each unigram and sift through each tweet. Therefore, as stated previously that is why only 5000 tweets per unigram and not more were used in the early stages of this PhD study. Another rare issue not necessarily a limitation, would be that this study is new and has not been studied before in this manner.

Moreover, it is important to include within the thesis the time frame that the data was collected from May-August 2020 during the global pandemic (coronavirus). In fact, perhaps the pandemic, lockdown, or isolation during this time could have contributed to the findings. Meaning were individuals' constantly on their computers since "real world life" was at a standstill? Giving this study or research a vital chance to collect a vast amount of data on the topic at hand. However, it is still possible that these findings would have been the same if coronavirus was not present because, the use of computers are involved in everyday life.

Selected limitations of the current methods are based on how open and immeasurable the Internet is, meaning that whoever is doing research on social media the opportunities are endless. Consequence to this, is the implication that can follow are infinite, if there are not set guidelines in place the researcher can gather data that is not efficient enough and is not used. However, with the Internet and social media are great tools to further the research but having a plain or guideline in place helps gather the data and not waste time on collecting data that is not used or needed. Due to the events of Corina Virus, certain tools and structures caused the study to be hinder slightly due to not having access to them during that time.

Another limitation that was unfortunate but was brought forth was the coast of certain programmes that the researcher wanted to use. Since the researcher had a personal budget for his or her project, he or she only used accessible or free programmes that were available to him or her. another algorithm and rewrite the code in numeric values and see the results that would come from that. From doing that rather than added more into this research that is on hand, it would be fascinating to see how either the data changes and the outcome varies, or the data does not change, and the outcome is in fact like what the chapter already details. All the infinite

possibilities or outcomes that this study or research can achieve is empirical to academia and cybercrime and cyberstalking detection. Which is needs to be talked about in a time or society like today, this research will open and pave way to the ongoing cybercrime detection activity and how it can be detected or flagged using data mining and machine learning.

7.3.2 Research Recommendation or Technique

The research recommendations or different technique for this study could be how the study is conducted wither with what different methods could be used or introducing a different dataset. For instance, the different variables that could be added or used are boundless whether it is using other algorithm(s) to compare the results obtained to one another if there is enough time to do so. Also, with have more time to conduct a study or research is greatly beneficial, however not always at the researcher's permission for that factor. As well as, using different programmes compared to NVivo and R Programming to see if there are differences or similarities.

7.4 Future Research

In the future the author would like to continue to construct a theoretical cyberstalking detection model (either with this study or with new data) that can perform data mining along with security metrics to detect cyberstalking from social media platforms with the use of Twitter. Also, to further the preliminary dataset and its outcomes. For instances, there was an option to include unigrams with hashtags within the collection of data; however, that could skew the data collected slightly, because the data collected would be closely related to the hashtag and unigram being used, not focusing on the links to cyberstalking and the unigram itself and not the hashtag. The different variables that could be added or used are boundless whether it is using

hashtags within Twitter, changing unigrams, focusing on the unigrams in more depth, or using a different from of social media all together. It would be very Interested in continuing this work or research by possibly running the same test and data with the opportunity to only use hashtag searches within Twitter. From doing that rather than added more into this research that is on hand, it would be fascinating to see how either the data changes and the outcome varies, or the data does not change, and the outcome is in fact like what the overall study already details.

Also, the author would like to consider reusing the data and software used for the preliminary dataset. As stated throughout this entire thesis the data mining methods that were used to measure parameters that the author was looking for. One of those parameters was location of postings or connections (IP addresses). Again, the location and postings or connections (IP address), is standard for this type of data mining machine learning software. Moreover, since all the information collected within that area was not used or needed for this study. It would be interesting to see how the IP addresses add to this study. To have more of an informal view of where around the world the cyberstalking indicative content tends to be the most prominent or even the areas where it is not and compare the similarities or differences. Furthermore, since KNN is very modest to implement and is most widely used as a first step in any machine learning setup. It is often used as a benchmark for more complex classifiers such as Support Vector Machines (SVM).

Henceforth, why the author would also like to use a different algorithm as well one algorithm that the author would like to conduct this study with is SVM. SVM is a supervised machine learning algorithm which can be used for classification or regression problems. It uses a technique called the kernel trick to transform data and then based on these transformations it finds an optimal boundary between the possible

outputs. However, it is mainly used on a numeric dataset, which the one used for this study was characterised. It would be interesting to see the results from SVM and compare them to KNN since both algorithms are used on a numeric dataset.

All the infinite possibilities or outcomes that this study or research can achieve is empirical to academia and cybercrime detection, this research will open and pave way to the ongoing cybercrime detection activity and how it can be detected or flagged using data mining and machine learning. A social structure of individuals related directly or indirectly based on some common factor like similar likings or retweets, is a social network. To understand the behaviour and structure of a social network it needs to study the network and this study is called social network analysis. Moreover, with all the unlimited opportunities or results that this study/ research can achieve is empirical to academia and cybercrime and cyberstalking detection, this research and its future work will continue to open and pave way to the ongoing cybercrime detection activity and how it can be detected or flagged using data mining and machine learning.

Appendix

Table 1TCyberstalking (no # used): 100 tweets via the most used terms or words.

Word	Count	Weighted Percentage	Similar Words
cyberstalking	1568	3.98%	#cyberstalkers, #cyberstalking, cyberstalked, cyberstalker, cyberstalkers, cyberstalking, cyberstalking'
harassment	451	1.15%	#harassment, harass, harassed, harasses, harassing, harassment
online	325	0.83%	#online, online
#metoo	324	0.82%	#metoo
data	268	0.68%	data
story	259	0.66%	stories, story
charge	253	0.64%	charge, charged, charges, charging
victim	227	0.58%	victim, victims
criminal	197	0.50%	criminal, criminality, criminally, criminals
sexual	192	0.49%	sexual, sexually
world	192	0.49%	world
actions	187	0.48%	action, actions

predatory	186	0.47%	predatory
thriller	186	0.47%	thriller
call	172	0.44%	call, called, calling, calls
violence	169	0.43%	violence
working	168	0.43%	work, worked, working, works
technology	144	0.37%	technology
surveillance	136	0.35%	#surveillance, surveillance
gender	136	0.35%	gender
abstracts	134	0.34%	abstracts
analyst	134	0.34%	analyst
people	132	0.34%	people, peoples, peoples', peoples’
allegedly	128	0.33%	allegation, allegations, allegedly
stalking	125	0.32%	#stalking, stalk, stalked, stalking
creating	124	0.32%	create, created, creating
arrested	111	0.28%	arrest, arrested
pushed	107	0.27%	push, pushed
hostile	97	0.25%	hostile
man	95	0.24%	man
porn	95	0.24%	porn

threats	94	0.24%	threat, threats
despite	94	0.24%	despite
revenge	94	0.24%	revenge
justify	91	0.23%	justify
defends	85	0.22%	defendant, defending, defends
cyber	82	0.21%	#cyber, cyber
police	75	0.19%	police
crimes	72	0.18%	#crime, #crimes, crime, crimes
media	70	0.18%	media
angry	69	0.18%	angry
woman	69	0.18%	woman
student	67	0.17%	#students, student
law	66	0.17%	law, laws
received	66	0.17%	received, receives, receiving
cyberbullying	62	0.16%	#cyberbullying, cyberbully, cyberbullying
publicly	61	0.16%	public, publically, publication, publicly
prison	60	0.15%	#prison, prison, prisoner
video	56	0.14%	video, videos
attack	56	0.14%	attack, attacker, attacking, attacks

like	52	0.13%	like, liked
engaging	49	0.12%	engage, engages, engaging
claims	48	0.12%	claim, claimed, claiming, claims
reporting	44	0.11%	report, reported, reportedly, reporter, reporters, reporting, reports
terrorists	44	0.11%	terroristic, terrorists
support	41	0.10%	support, supported, supporter, supporters, supporting
confinement	41	0.10%	confinement

Table 2 Stalking and Fear (no # used) 100 tweets via the most used terms or words.

Word	**Count**	**Weighted Percentage**	**Similar Words**
stalks	17635	6.47%	#stalk, #stalked, #stalking, 'stalking, 'stalking', @stalking, stalk, stalked, stalking, stalking', stalking’, stalks
liking	2010	0.74%	like, liked, likely, likes, liking
stops	1256	0.46%	stop, stopped, stopping, stops

accounts	1151	0.42%	account, accountability, accountancy, accounts
tweets	939	0.34%	tweet, tweeted, tweeting, tweets, tweets'
peoples	937	0.34%	@people, people, peoples
ones	836	0.31%	#one, one, ones
videos	737	0.27%	video, videos
kills	712	0.26%	#killing, 'killing, kill, killed, killing, kills
twitters	701	0.26%	#twitter, @twitter, twitter, twitters
harassment	688	0.25%	harass, harassed, harassement, harasser, harassers, harasses, harassing, harassment, harassments
timing	660	0.24%	@time, time, times, time', timing
rights	656	0.24%	right, rights
trends	612	0.22%	trend, trending, trends
using	555	0.20%	use, useful, uses, using
looks	544	0.20%	look, looked, looking, looks

finds	542	0.20%	#find, find, finding, findings, finds
instagrams	538	0.20%	#instagram, instagram, instagrams
bitching	491	0.18%	bitch, bitches, bitching
someones	482	0.18%	someone, someones
follows	473	0.17%	follow, followed, follower, followers, following, followings, follows
think	463	0.17%	think, thinking, thinks
attention	444	0.16%	attention
girls	442	0.16%	@girls, girl, girls, girls'
wants	412	0.15%	#wanted, want, wanted, wanting, wants
something	407	0.15%	something
posts	388	0.14%	post, poste, posted, posting, posts
watching	367	0.13%	watch, watched, watches, watchful, watching
friends	356	0.13%	'friend', friend, friendly, friends, friends', friends’, friend’

calls	355	0.13%	@call, call, called, calling, calls
feels	355	0.13%	#feelings, feel, feeling, feelings, feels
abusive	336	0.12%	#abuse, abuse, abuse', abused, abuser, abusers, abuses, abusing, abusive
persons	333	0.12%	'person', person, personal, personalities, personality, personally, persone, persons
threats	319	0.12%	threat, threats
views	318	0.12%	view, viewed, viewing, views
talking	309	0.11%	talk, talked, talking, talks
socials	303	0.11%	social, sociale, socially, socials
hashtag	300	0.11%	hashtag, hashtags
tells	299	0.11%	tell, telling, tells
responsible	292	0.11%	response, responses, responsibilities, responsibility, responsible

everyone	287	0.11%	everyone, everyones
trying	283	0.10%	@try, tried, tries, try, trying
family	283	0.10%	#family, families, family
guys'	280	0.10%	guy, guys, guys'
medias	272	0.10%	#media, media, medias
police	264	0.10%	@police, police, policing
crazy	262	0.10%	crazies, craziness, crazy
hating	262	0.10%	hate, hated, hateful, hates, hating

Table 3

Table 3 #Cyberstalking 100 tweets via the most used terms or words out of 150

Word	**Length**	**Count**	**Weighted Percentage**	**Similar Words**
cyberstalking	13	222	3.97%	#cyberstalkers, #cyberstalking, cyberstalker, cyberstalking
violence	8	126	2.25%	violence
#metoo	6	122	2.18%	#metoo
technology	10	122	2.18%	technology
call	4	121	2.16%	call, calls
#digitalactivism	16	120	2.14%	#digitalactivism
#imagebasedsexualabuse	22	120	2.14%	#imagebasedsexualabuse
#onlinemisogyny	15	120	2.14%	#onlinemisogyny
#surveillance	13	120	2.14%	#surveillance
#technologyfacilitatedabuse	27	120	2.14%	#technologyfacilitatedabuse
gender	6	120	2.14%	gender
stalking	8	45	0.80%	#stalking, stalking
online	6	32	0.57%	#online, online

#cyberbullying	14	25	0.45%	#cyberbullying, cyberbullying
harass	6	15	0.27%	#harassment, harass, harasses, harassment
cyber	5	14	0.25%	#cyber, cyber
#cybercrime	11	14	0.25%	#cybercrime, cybercrime
#cyberharassment	16	13	0.23%	#cyberharassment
#cybersafety	12	13	0.23%	#cybersafety
#cybersecurity	14	12	0.21%	#cybersecurity
security	8	12	0.21%	security
victims	7	9	0.16%	victim, victims
#women	6	8	0.14%	#women, women
abusive	7	8	0.14%	abuse, abusive
individuals	11	2	0.04%	individuals
messages	8	2	0.04%	messages, messaging
monitor	7	2	0.04%	monitor, monitoring
threat	6	2	0.04%	threat, threats
#cyberbullyng	13	1	0.02%	#cyberbullyng
#dangersofcyberstalkers	23	1	0.02%	#dangersofcyberstalkers
fear	4	1	0.02%	fear
find	4	1	0.02%	find
friends	7	1	0.02%	friends
girl	4	1	0.02%	girl
guys	4	1	0.02%	guys
hide	4	1	0.02%	hide
like	4	1	0.02%	like
media	5	1	0.02%	media
sexting	7	1	0.02%	sexting
sharing	7	1	0.02%	sharing
sinister	8	1	0.02%	sinister
social	6	1	0.02%	social
source	6	1	0.02%	source
superior	8	1	0.02%	superior
surviving	9	1	0.02%	surviving
targets	7	1	0.02%	targets

Table 4

Table 1. Word Frequency out of 50

Word	Length	Count	Weighted Percentage	Similar Words
unwanted’	9	50425	0.71%	#unwant, #unwanted, "unwanted"faces, 'unwanted, 'unwanted', @unwanted, unwanted, unwanted', unwanted’
stalker’iyla	12	43359	0.61%	#stalker, #stalkers, '#stalker, "stalker, "stalker", 'stalker, 'stalker', @stalker, stalker, stalker#, stalker##, stalker', stalkers, stalker’, stalker’da, stalker’iyla, stalker’ım, stalker’’, stalker—
messaging’	10	43000	0.61%	#message, #messaging, 'messaging', @message, message, messaged, messages, messages', messages’, message’, messaging, messaging', messaging’
stalks	6	37318	0.53%	#stalk, #stalking, 'stalk, 'stalk', 'stalking, 'stalking', @stalk, @stalking, stalk', stalke, stalked, stalking, stalking', stalking’, stalking’e, stalks, stalk’, stalk’n, stalk’r, stalk’u
people’	7	36797	0.52%	#people, "people, 'people, @people, @peoples, people, people', peoples, peoples', peoples’, people’
□follow	7	27507	0.39%	##follow, #follow, #followal, #followed, #follower, #followers, #following, #follows, #follows#followgram, 'follow, 'follows, @follow, follow, follow', followed, follower, followering, followers, followers', following, followings, following’, followment, follows, follow’, follow, follow, follow, □follower
harassment’	11	21825	0.31%	#harass, #harasser, #harassing, #harassment, 'harassing', 'harassment, 'harassment', harass, harassed, harassement, harasser, harassers, harasses, harassing, harassment, harassment', harassments, harassment’
feelings	8	21664	0.30%	#feeling, 'feel', @feeling, feeling, feelings, feelings’
removing	8	20975	0.30%	#removable, #removal, #remove, removable, removal, removals, remove, removed, remover, removes, removing, removings

rejects	7	19799	0.28%	#rejection, 'rejects', reject, rejected, rejecting, rejection, rejections, rejects
annoys	6	19127	0.27%	#annoying, 'annoying, 'annoying', @annoying, annoyance, annoyances, annoyed, annoying, annoying', annoyingly, annoying', annoys
scaring	7	18226	0.26%	#scare, #scared, 'scared', @scared, scared, scared', scared', scares, scaring
technology'	11	16558	0.23%	#technologies, #technology, 'technology, @technology, technologic, technological, technologically, technologies, technologies', technology, technology', technology'
pleasing	8	14508	0.20%	please, pleased, pleases, pleasing, please
violent'	8	14389	0.20%	#violent, 'violent, 'violent', @violent, violent, violent', violente, violentes, violently, violents, violent'
friend'	7	14269	0.20%	#friend, #friendly, #friends, 'friend', 'friends, 'friends', @friend, @friends, friend, friend', friended, friending, friendlies, friendly, friendly', friends, friends', friends', friend'
victim'	7	14083	0.20%	#victim, #victimized, #victims, 'victim, 'victim', 'victims', @victim, victim, victim#6, victim', victime, victimes, victimization, victimize, victimized, victimizer, victimizers, victimizes, victimizing, victims, victims', victims', victim', victimization
sexually	8	13576	0.19%	#sexual, 'sexual, 'sexually, @sexual, sexual, sexualities, sexuality, sexualization, sexualize, sexualized, sexualizes, sexualizing, sexuall, sexually
policing	8	12482	0.18%	#police, #policing, 'police, @police, police, police', policed, policer, polices, police', policing
someones	8	10483	0.15%	someon, someone, someones
twitters	8	10362	0.15%	#twitter, 'twitter', @twitter, twitter, twitterer, twitters, twitter'a

persons	7	9971	0.14%	#person, #personal, #personality, "personal, 'person, @person, @personal, person, person', personable, personal, personalities, personality, personality', personalization, personalizations, personalize, personalized, personalizing, personally, personals, personation, persone, persons, person'
really	6	9307	0.13%	really
calling	7	8973	0.13%	#calling, #calls, 'call', call#npc, called, called', called', calling, calls'
□report	7	8366	0.12%	#report, #reporting, 'report', @reporter, @reports, report, report', reported, reportedly, reporter, reporters, reporter', reporting, reporting', reportings, reports, report, □reporting
accounts	8	8287	0.12%	#account, #accountability, #accountable, #accountant, #accountants, #accounting, #accounts, @account, account, accountability, accountable, accountancy, accountant, accountants, accounted, accounting, accounts, accounts'
thinks	6	8096	0.11%	#think, #thinking, @think, @thinks, think', thinking, thinks
portlanders	11	8038	0.11%	#portland, 'portland', portland, portlander, portlanders
without	7	7296	0.10%	without, without
public'	7	6972	0.10%	#public, 'public, 'publication, 'publicly', @public, public, publically, publication, publications, publicity, publicize, publicized, publicly, publics, public'
creepy'	7	6932	0.10%	#creepy, 'creepy, 'creepy', @creepy, creepiness, creepy, creepy'
president	9	6893	0.10%	#presidency, presided, presidency, president, presidenting, presidents, presidents', presiding

students’	9	6794	0.10%	#student, #students, student, student', students, students', students’
supports	8	6579	0.09%	#support, 'support', @support, support, supportable, supportant, supporte, supported, supportent, supporter, supporters, supporting, supportive, supports
□protect	8	6567	0.09%	#protect, #protection, "protect", 'protecting', @protect, protect, protected, protected’, protecter, protecting, protection, protections, protective, protectively, protects, protect’, protect
@realdonaldtrump	16	6275	0.09%	@realdonaldtrump, realdonaldtrump
always	6	6252	0.09%	#always, 'always, @always, always, always’
creepers	8	6213	0.09%	#creeper, #creepers, 'creeper, 'creeper', @creeper, @creepers, creeper, creeper#, creeper#2867, creeper', creepers
trying	6	6115	0.09%	trying, trying
federalism	10	6007	0.08%	#federal, @federal, federal, federalism, federally, federated, federation, federations
country'	8	5939	0.08%	#country, @country, countries, countries’, country, country', countrys, country’
reason’	7	5770	0.08%	#reason, 'reasonable, @reason, reason, reason#384837717171, reasonable, reasonableness, reasonably, reasoned, reasoning, reasonings, reasoning’, reasons, reason’
schools	7	5742	0.08%	#school, #schools, school, school', schooling, schools
disgusts	8	5668	0.08%	#disgust, #disgusting, 'disgusting', disgust, disgusted, disgusting, disgustingly, disgusts
distancing	10	5419	0.08%	distance, distanced, distances, distancing

everyones	9	5348	0.08%	everyone, @everyone, everyone, everyone', everyones
abusively	9	5204	0.07%	#abuse, #abused, #abuser, #abusers, #abusive, 'abuse', 'abuser', @abuse, abuse', abused, abusent, abuser, abuser'', abusers, abuses, abuse’, abusing, abusive, abusively, abusiveness, abusive’, abuse
seconds	7	5099	0.07%	second, seconde, seconded, secondes, secondly, seconds
domestics	9	5053	0.07%	#domestic, 'domestic, domestic, domestically, domesticated, domestics
secrets	7	5033	0.07%	#secrets, 'secret, 'secret', 'secretive, secret, secret', secrete, secretely, secretions, secretive, secretly, secrets

Table 5

Table 2 Stalking and Fear (no # used): 100 tweets via the most used terms or words out of 150.

Word	**Count**	**Weighted Percentage**	**Similar Words**
stalks	17635	6.47%	#stalk, #stalked, #stalking, 'stalking, 'stalking', @stalking, stalk, stalked, stalking, stalking', stalking’, stalks
liking	2010	0.74%	like, liked, likely, likes, liking
stops	1256	0.46%	stop, stopped, stopping, stops

accounts	1151	0.42%	account, accountability, accountancy, accounts
tweets	939	0.34%	tweet, tweeted, tweeting, tweets, tweets'
peoples	937	0.34%	@people, people, peoples
ones	836	0.31%	#one, one, ones
videos	737	0.27%	video, videos
kills	712	0.26%	#killing, 'killing, kill, killed, killing, kills
twitters	701	0.26%	#twitter, @twitter, twitter, twitters
harassment	688	0.25%	harass, harassed, harassement, harasser, harassers, harasses, harassing, harassment, harassments
timing	660	0.24%	@time, time, times, time', timing
rights	656	0.24%	right, rights
trends	612	0.22%	trend, trending, trends
using	555	0.20%	use, useful, uses, using
looks	544	0.20%	look, looked, looking, looks

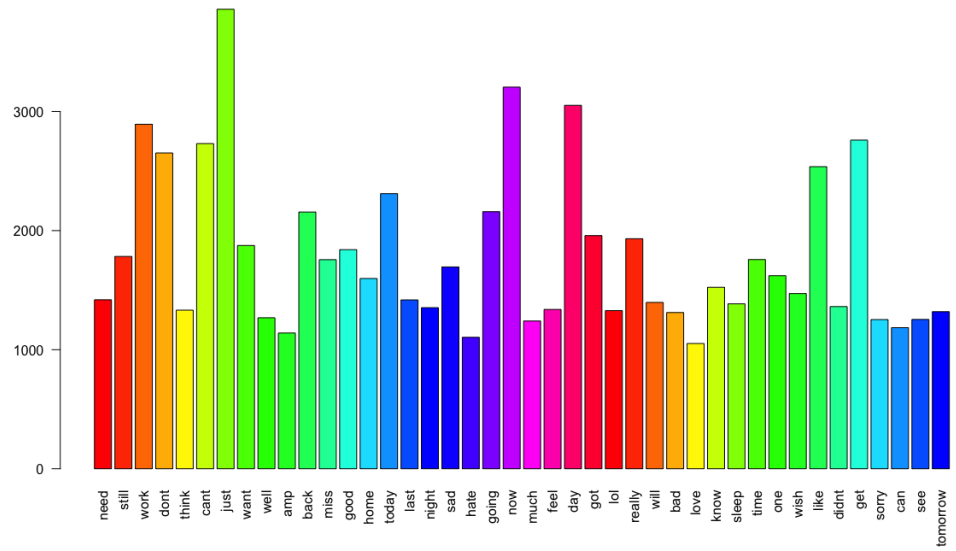
finds	542	0.20%	#find, find, finding, findings, finds
instagrams	538	0.20%	#instagram, instagram, instagrams
bitching	491	0.18%	bitch, bitches, bitching
someones	482	0.18%	someone, someones
follows	473	0.17%	follow, followed, follower, followers, following, followings, follows
think	463	0.17%	think, thinking, thinks
attention	444	0.16%	attention
girls	442	0.16%	@girls, girl, girls, girls'
wants	412	0.15%	#wanted, want, wanted, wanting, wants
something	407	0.15%	something
posts	388	0.14%	post, poste, posted, posting, posts
watching	367	0.13%	watch, watched, watches, watchful, watching
friends	356	0.13%	'friend', friend, friendly, friends, friends', friends’, friend’

calls	355	0.13%	@call, call, called, calling, calls
feels	355	0.13%	#feelings, feel, feeling, feelings, feels
abusive	336	0.12%	#abuse, abuse, abuse', abused, abuser, abusers, abuses, abusing, abusive
persons	333	0.12%	'person', person, personal, personalities, personality, personally, persone, persons
threats	319	0.12%	threat, threats
views	318	0.12%	view, viewed, viewing, views
talking	309	0.11%	talk, talked, talking, talks
socials	303	0.11%	social, sociale, socially, socials
hashtag	300	0.11%	hashtag, hashtags
tells	299	0.11%	tell, telling, tells
responsible	292	0.11%	response, responses, responsibilities, responsibility, responsible

everyone	287	0.11%	everyone, everyones
trying	283	0.10%	@try, tried, tries, try, trying
family	283	0.10%	#family, families, family
guys'	280	0.10%	guy, guys, guys'
medias	272	0.10%	#media, media, medias
police	264	0.10%	@police, police, policing
crazy	262	0.10%	crazies, craziness, crazy
hating	262	0.10%	hate, hated, hateful, hates, hating

Bar-Plots and Rank Tables from Datasets 2-5

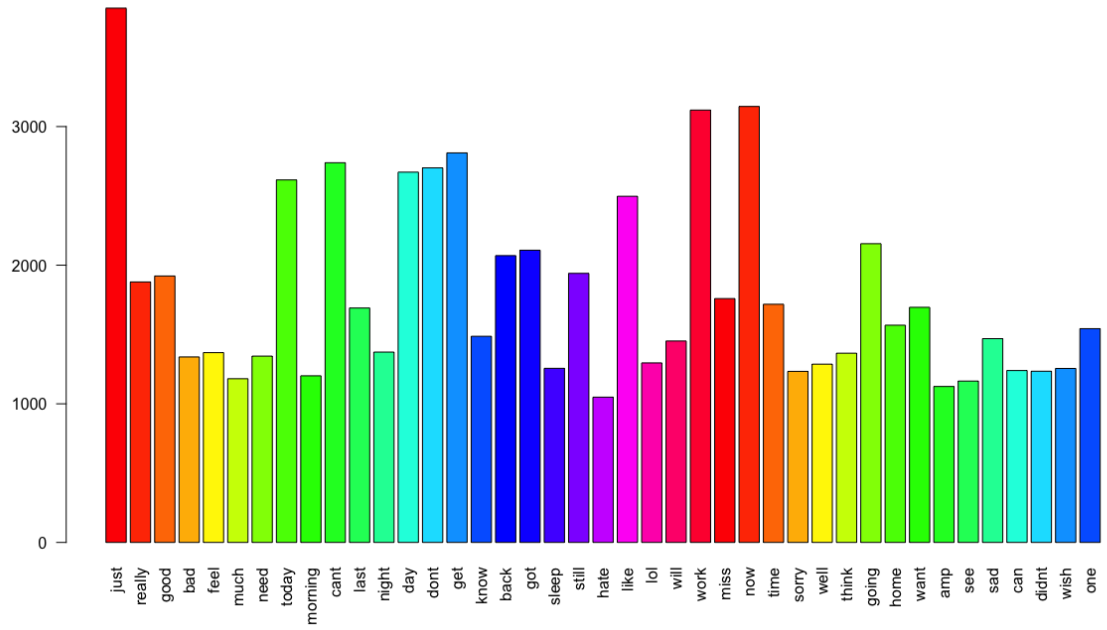
Figure 28. Bar-Plot Dataset 2 and Rank Table



Freq Word	Count	Rank
just	3859	1
now	3205	2
day	3052	3

work	2893	4
get	2760	5
cant	2731	6
don't	2652	7
like	2537	8
today	2310	9
going	2159	10
back	2156	11
got	1958	12
really	1932	13
want	1875	14
good	1840	15
still	1783	16
time	1757	17
miss	1756	18
sad	1695	19
one	1621	20
home	1598	21
know	1524	22
wish	1471	23
need	1419	24
last	1418	25
will	1397	26
sleep	1385	27
didn't	1362	28
night	1353	29
feel	1338	30
think	1332	31
lol	1328	32
tomorrow	1319	33
bad	1312	34
well	1268	35
see	1254	36
sorry	1253	37
much	1241	38
amp	1185	39
hate	1104	40
love	1052	41

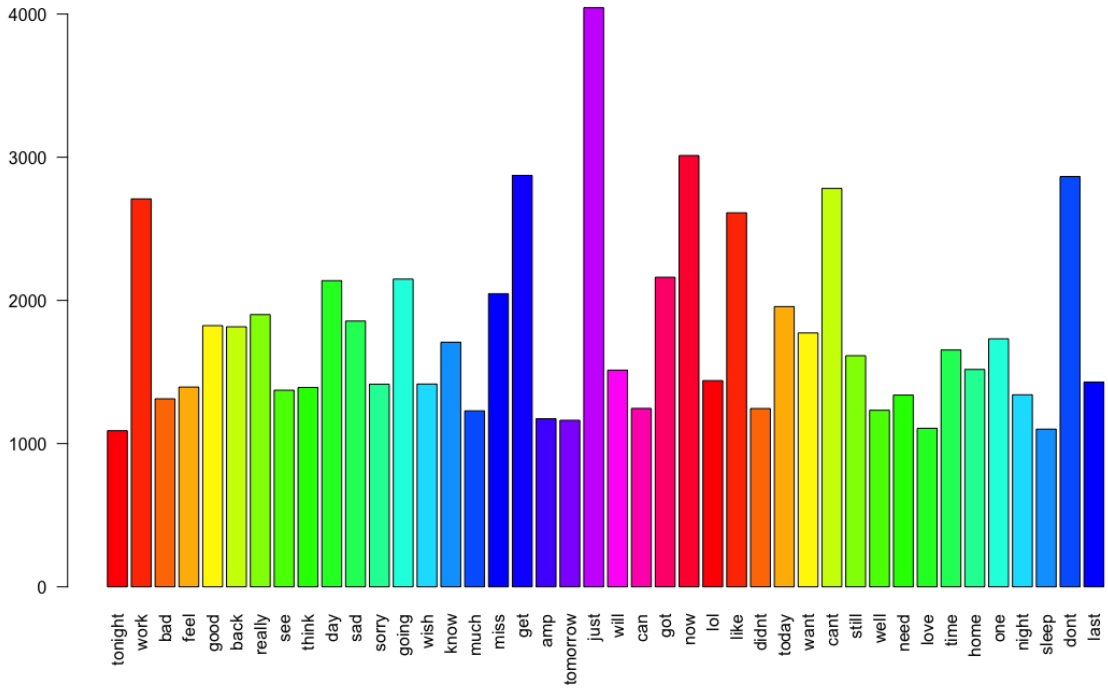
Figure 29. Bar-Plot Dataset 1 and Rank Table.



Word Freq	Count	Rank
just	3854	1
now	3145	2
work	3119	3
get	2810	4
cant	2739	5
don't	2702	6
day	2671	7
today	2615	8
like	2497	9
going	2155	10
got	2108	11
back	2069	12
still	1941	13
good	1922	14
really	1879	15
time	1718	17
miss	1759	16
want	1691	18
last	1691	18
home	1576	20
one	1542	21
know	1487	22
sad	1470	23
will	1453	24

night	1373	25
feel	1369	26
think	1365	27
need	1344	28
bad	1338	29
lol	1295	30
well	1286	31
sleep	1256	32
wish	1255	33
can	1240	34
didn't	1235	35
sorry	1234	36
morning	1202	37
much	1181	38
see	1164	39
amp	1125	40
hate	1048	41

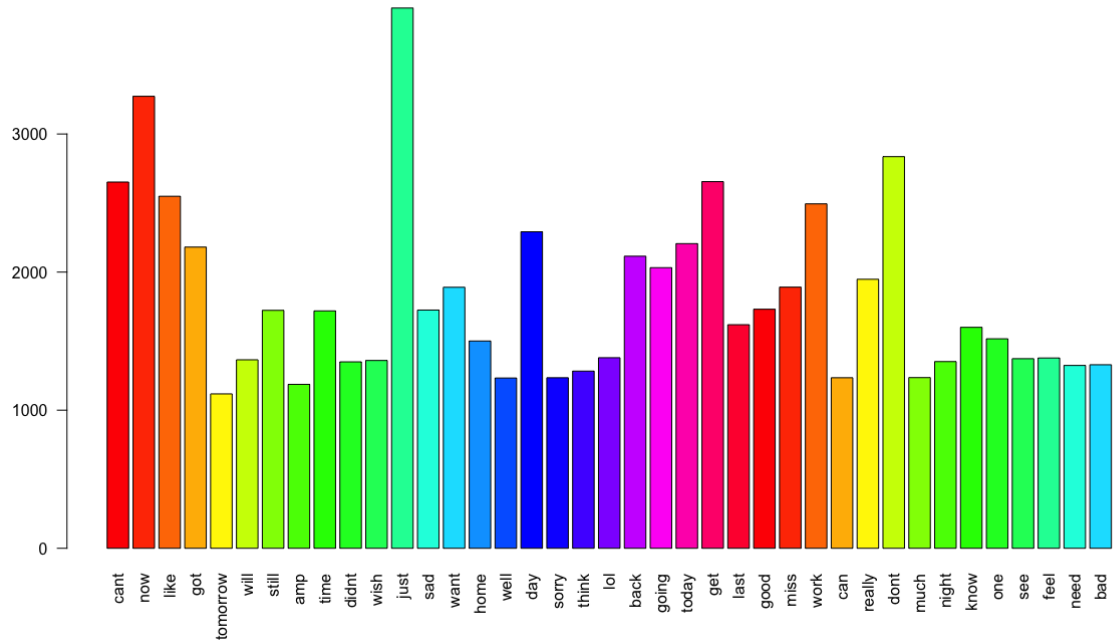
Figure 30. Bar-Plot Dataset 2 and Rank Table.



Word Freq	Count	Rank

just	4004	1
now	3012	2
get	2873	5
don’t	2895	4
cant	2782	6
work	2709	7
like	2612	8
got	2162	9
going	2149	10
day	2138	11
miss	2947	3
today	1957	12
really	1901	13
sad	1856	14
good	1824	15
back	1816	16
wont	1773	17
one	1732	18
know	1708	19
time	1654	20
still	1614	21
home	1518	22
will	1513	24
lol	1440	25
last	1430	26
wish	1416	27
sorry	1415	28
feel	1395	29
think	1518	22
see	1373	30
night	1341	31
need	1339	32
bad	1313	33
can	1246	34
didn’t	1245	35
well	1233	36
much	1229	37
amp	1174	38
tomorrow	1163	39
love	1107	40
tonight	1090	41

Figure 40. Bar-Plot Dataset 4 and Rank Table.



Word Freq	Counr	Rank
just	3912	1
now	3273	2
don't	2836	3
get	2665	4
cant	2652	5
like	2549	6
work	2494	7
day	2292	8
tody	2206	9
got	2181	10
back	2115	11
going	2032	12
really	1948	13
miss	1891	14
want	1890	15
good	1731	16
sad	1725	17
still	1723	18
time	1719	19
last	1619	20

know	1600	21
one	1517	22
home	1501	23
lol	1380	24
feel	1378	25
see	1373	26
will	1365	27
wish	1360	28
night	1352	29
didn't	1350	30
bad	1329	31
need	1324	32
think	1282	33
much	1236	34
sorry	1235	35
can	1235	35
well	1233	37
amp	1187	38
tomorrow	1118	39

Figure 41. Bar-Plot Dataset 5 and Rank Table.

Advanced Search in RStudio: csv files datasets 2-5

Dataset 2.csv (6 Tweets)

Abuse, Annoying, Creep/Creepy, Fear, Gender, Harassment, Messaging, Relationships, Scared, Unwanted, Victim, Violent = 0

Follow/Follows 1

> grep("follow", tweets$text)
[40887]
> tweets$text [40887]
[1] are spammers starting to spam tweet in addition to just following ppl? that surely spells the end of twitter, u know?

Stalker: 3

> grep("stalker", tweets$text)
[5091 14223 20749]

> tweets$text [5081]

[1] [redacted] ah yes, the joy of cyberstalkers. sorry about that.

> tweets$text [14223]
[1] [redacted] you think i'm do not even come close to a subscription of a stalker!

> tweets$text [20749]
[1] [redacted] yay for stalking indeed.

Technology: 2
grep("technology", tweets$text)
[43016 7412]

> tweets$text [43016]
[1] I dont think anyone understand the pain this is causing me I hate technology right now. fuck u twitter!

> tweets$text [7412]
[1] How did tweetdeck log into my facebook w/o me giving it my FB credentials.? Ayo technology....

Dataset 3.csv (11 Tweets)

Annoying, Fear, Gender, Harassment, Messaging, Relationships, Scared, Technology, Unwanted, Victim, Violent = 0

Abuse 1
> grep("abuse", tweets$text)

> tweets$text [45941]
[1] got cyber bullied by an adult online, CHILD ABUSE. poor poor bay, shes not allowed to go to the millpond now

Creep/ Creepy 4
>grep("creep", tweets$text)
[20398 20415 25740 42782]

> tweets$text [20387]
[1] @compwallpaper YOU BIG CREEP LEAVE ME ALONE!

> tweets$text [20415]
[1] I DONT WANT YOU TO TALK ME TO ME YOU CREEP ME OUT STALKER MAKE ME SAD (stalker)

> tweets$text [42782]

[1] [redacted] ahahaa! I hate when creepers try to follow me. I've had the same girl try to follow me three times in like 2 days. Sick

> tweets$text [25740]
[1] Maybe it's all true, I'm a disgusting creepy stalker old cat lady in the making and I shouldn't be allowed on the internet.

Follow/Follows 1
> grep("follow", tweets$text)
[40887]
> tweets$text [40887]
[1] are spammers starting to spam tweet in addition to just following ppl? that surely spells the end of twitter, u know?

Stalker: 5
> grep("stalker", tweets$text)
[18361 20414 1489 33346 46984]

> tweets$text [20415]
[1] I DONT WANT YOU TO TALK ME TO ME YOU CREEP ME OUT STALKER MAKE ME SAD (creep)

> tweets$text [1489]
[1] [redacted] dude this twitter shit is freakin scary...I fuckin Located your ass ahahahaha right in NEU...I hope none of my stalkers find me !

> tweets$text [33346]
[1] [redacted] oh the joys of stalking people....seriously it just becomes easier and easier.

> tweets$text [46984]
[1] spammers are stalking me now on twitter

> tweets$text [18361]
[1] [redacted] wow really keep writing letters and stalking her. she'll give in!!

Dataset 4.csv (10 Tweets)

Abuse, Fear, Gender, Harassment, Messaging, Relationships, Technology, Unwanted, Victim, Violent = 0

Annoying 1
> grep("annoying", tweets$text)
[36907]

> tweets$text [36907]
[1] [redacted] Hi! This is your annoying-ass stalker from Costa Rica! Remember? Cincinnati? Hitchens too! Am I less annoying? No (annoying)

Creep/Creepy 4
> grep("creepy", tweets$text)
[11186 15190 33157 39436]

> tweets$text [11186]
[1] Ew I have creepy followers

> tweets$text [15190]
[1] I AM SUCH A CREEPER I feel disappointed because of it. Damn my cyberstalking skills the internet = no more privacy.

> tweets$text [33157]
[1] [redacted] You've got to be kidding me! Twitter makes me feel like a creepy ass stalker.....(stalker)

> tweets$text [39436]
[1] wondering how these freaks always seem to find me? so creepy...I mean seriously is their something that bad wrong with me?

Follow/Follows 1
> grep("follow", tweets$text)
[111868]

> tweets$text [11186]
[1] Ew I have creepy followers

Scared: 1
> grep("scared", tweets$text)
[15069]

> tweets$text [15069]
[1] oh no somebody hacked into my email i'm scared now. what the fuck!!

Stalker: 3
> grep("stalker", tweets$text)
[22573 33157 36907]

> tweets$text [25573]
[1] The only way I'll feel safe is when he's finally in jail where his stalker ass belongs!

> tweets$text [33157]
[1] [redacted] You've got to be kidding me! Twitter makes me feel like a creepy ass stalker.....(stalker)

> tweets$text [36907]
[1] [redacted] Hi! This is your annoying-ass stalker from Costa Rica! Remember? Cincinnati? Hitchens too! Am I less annoying? No (annoying)

Dataset 5.csv (7 Tweets)

Annoying, Fear, Follow/Follows, Gender, Harassment, Messaging, Relationships, Scared, Technology, Unwanted, Victim, Violent = 0

Abuse 2
> grep("abuse", tweets$text)
[40267 15992]

> tweets$text [40267]
[1] i seem to get a lot of abuse from twitter users, am considering leaving because of it... strangers who dont know me seem to think they do

> tweets$text [15992]
[1] @_MattyJones You are just horrible :L... abuse over twitter again! I have a right to just block you (;(;

Creep/Creepy 2
> grep("creepy", tweets$text)
> grep("creep", tweets$text)
[2819 20910]

> tweets$text [2819]
[1] There are some really creepy men on Trekspace. and they love to hit on little ol' me.

> tweets$text [20910]
[1] [redacted] Y didn't u tell me u made a twitter account instead of just follow me like a creepo!!!

Stalker: 3
> grep("stalker", tweets$text)
[14459 27126 40891]

> tweets$text[14459]
[1] [redacted] I don't know how these people find me either.. Creepy stalker types

> tweets$text [27126]
[1] [redacted] Oh my goodness. How very scary...a stalker.

> tweets$text [40891]
[1] I have a stalker...for real. I won't be allowing anonymous comments on my blog for a while...

References

A. A. Moore, “Cyberstalking and Women: Facts and Statistics,” thoughtco.com, 2018.

Abbass, Z., Ali, Z., Ali, M., Akbar, B. and Saleem, A., 2020, February. A Framework to Predict Social Crime through Twitter Tweets by Using Machine Learning. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)* (pp. 363-368). IEEE.

Adedoyin-Olowe, M., Gaber, M.M., Dancausa, C.M. and Stahl, F., 2014, December. Extraction of unexpected rules from twitter hashtags and its application to sport events. In *2014 13th International Conference on Machine Learning and Applications* (pp. 207-212). IEEE.

Aggarwal, C.: An introduction to social network data analytics. Springer US, 2011.

Alexander, B., 2006. Web 2.0: A new wave of innovation for teaching and learning?. *Educause review*, *41*(2), p.32.

Alves, F., Bettini, A., Ferreira, P.M. and Bessani, A., 2021. Processing tweets for cybersecurity threat awareness. *Information Systems*, *95*, p.101586.

Amato, F., Boselli, R., Cesarini, M., Mercorio, F., Mezzanzanica, M., Moscato, V., Persia, F. and Picariello, A., 2015, February. Challenge: Processing web texts for classifying job offers. In *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)* (pp. 460-463). IEEE.

Anderson, R., Barton, C., Böhme, R., Clayton, R., Van Eeten, M.J., Levi, M., Moore, T. and Savage, S., 2013. *Measuring the cost of cybercrime*. In The economics of information security and privacy (pp. 265-300). Springer, Berlin, Heidelberg.
Vancouver.

Ang, S. and Van Dyne, L., 2015. Conceptualization of cultural intelligence: Definition, distinctiveness, and nomological network. *In Handbook of cultural intelligence* (pp. 21-33). Routledge.

A. Sechelea, T. Do Huu, E. Zimos, and N. Deligiannis, “Twitter data clustering and visualization,” in Telecommunications (ICT), 2016 23rd International Conference on. IEEE, 16 May 2016, pp. 1–5.

Analytics R Programming Guide , UC Business. “K-Means Cluster Analysis.” *K-Means Cluster Analysis · UC Business Analytics R Programming Guide*, 2 Oct. 2018, uc-r.github.io/kmeans_clustering.

Authors/Task Force members, Elliott, P.M., Anastasakis, A., Borger, M.A., Borggrefe, M., Cecchi, F., Charron, P., Hagege, A.A., Lafont, A., Limongelli, G. and Mahrholdt, H., 2014. 2014 ESC Guidelines on diagnosis and management of hypertrophic cardiomyopathy: the Task Force for the Diagnosis and Management of Hypertrophic Cardiomyopathy of the European Society of Cardiology (ESC). *European heart journal*, *35*(39), pp.2733-2779.

Baars, H., & Kemper, H.-G. (2008). Management support with structured and un- structured data – An integrated business intelligence framework. Information Systems Management, 25(2), 132–148. http://dx.doi.org/10.1080/10580530801941058.

Barclay, C., Donalds, C. and Osei-Bryson, K.M., 2018. Investigating critical success factors in online learning environments in higher education systems in the Caribbean. *Information Technology for Development*, *24*(3), pp.582-611.

Beier, M., & Wagner, K. (2016). Social media adoption: barriers to the strategic use of social media in SMEs. *Proceedings of the european conference on information systems.*

Bilal, M., Hussain, A., Jaffar, M.A., Choi, T.S., Mirza, A.M., Alamri, A., Hossain, M.S., Almogren, A., Hassan, M.M., Alnafjan, K. and Zakariah, M*., Quality of Service Aware Reliable Task Scheduling in Vehicular Cloud Computing*. Mob. Netw. Appl., 372, pp. 2045-2053.

Bijalwan, V., Kumar, V., Kumari, P. and Pascual, J., 2014. KNN based machine learning approach for text and document mining. *International Journal of Database Theory and Application*, *7*(1), pp.61-70.

L. Bijuraj, “Clustering and its applications,” in Proceedings of National Conference on New Horizons in IT-NCNHIT, 2013, pp. 169-172.

Bocij, P. and McFarlane, L., 2003. Cyberstalking: The technology of hate. *The Police Journal*, *76*(3), pp.204-221.

Bonchi, F., Castillo, C., Gionis, A., and Jaimes, A. Social network analysis and mining for business applications. ACM Transactions on Intelligent Systems and Technology 2, 3 (Apr. 2011), 22.

Boyd, D.M. and Ellison, N.B., 2007. Social network sites: Definition, history, and scholarship. *Journal of computer-mediated Communication*, *13*(1), pp.210-230.

Buckland, S.T., Plumptre, A.J., Thomas, L. and Rexstad, E.A., 2010. Design and analysis of line transect surveys for primates. *International Journal of Primatology*, *31*(5), pp.833-847.

Burba, F., Ferraty, F. and Vieu, P., 2009. k-Nearest Neighbour method in functional nonparametric regression. *Journal of Nonparametric Statistics*, *21*(4), pp.453-469.

Bureau Of Justice Statistics, "Stalking". (2015) [Online]. Available from: 10 July 2019 https://www.bjs.gov/index.cfm?ty=tp&tid=973

Butkovic, A., Mrdovic, S., Uludag, S. and Tanovic, A., 2019. Geographic profiling for serial cybercrime investigation. *Digital Investigation*, *28*, pp.176-182.

Cadar, C. and Sen, K., 2013. Symbolic execution for software testing: three decades later. *Communications of the ACM*, *56*(2), pp.82-90.

Cambria, E., Rajagopal, D., Olsher, D. and Das, D., 2013. Big social data analysis. *Big data computing*, *13*, pp.401-414.

Casey, E., 2011. *Digital evidence and computer crime: Forensic science, computers, and the Internet*. Academic press.

Cassell, J. and Cramer, M., 2008. *High tech or high risk: Moral panics about girls online*. MacArthur Foundation Digital Media and Learning Initiative.

Castells, M., 2002. *The Internet galaxy: Reflections on the Internet, business, and society.* Oxford University Press on Demand.Vancouver

Carol M. Walker, Beth Rajan Sockman and Steven Koehn, 2011. An exploratory study of cyberbullying with undergraduate university students.

Cebeci, Z., & Yildiz, F. (2015). Comparison of k-means and fuzzy c-means algorithms on different cluster structures. Agrárinformatika/Journal of Agricultural Informatics, 6(3): 13–23. https://doi.org/10.17700/jai.2015.6.3.196

Chen, S., Guo, C., Yuan, X., Zhang, J., & Zhang, X. L. (2014). MovementFinder: *Visual analytics of origin-destination patterns from geo-tagged social media*. 2014 IEEE

conference on visual analytics science and technology, VAST 2014 – proceedings, 239–240. http://dx.doi.org/10.1109/VAST.2014.7042509.

Chen, X., Vorvoreanu, M., & Madhavan, K. P. C. (2014). *Mining social media data for understanding students' learning experiences.* IEEE Transactions on Learning Technologies, 7(3), 246–259. http://dx.doi.org/10.1109/TLT.2013.2296520.

Chen, S., Yuan, X., Wang, Z., Guo, C., Liang, J., Wang, Z., ... Zhang, J. (2016). *Interactive Visual Discovering of Movement Patterns from Sparsely Sampled Geo-tagged Social Media Data*. IEEE Transactions on Visualization and Computer Graphics, 22(1), 270–279. http://dx.doi.org/10.1109/TVCG.2015.2467619.

Cheng, D., Zhang, S., Deng, Z., Zhu, Y. and Zong, M., 2014, December. kNN algorithm with data-driven k value. In *International Conference on Advanced Data Mining and Applications* (pp. 499-512). Springer, Cham.

Choi, K.S., 2008. Computer crime victimization and integrated theory: An empirical assessment. *International Journal of Cyber Criminology*, *2*(1).

Coombs, Ted. Artificial Intelligence & Cybersecurity For Dummies®, IBM Limited Edition. *John Wiley and Sons, Inc,* 2018.

Cortizo, J., Carrero, F., Gomez, J., Monsalve, B., Puertas, E.: Introduction to Mining SM. In: *Proceedings of the 1st International Workshop on Mining SM*, 1 – 3, 2009.

CPS, Stalking and Harassment, Updated 23 May 2018, https://www.cps.gov.uk/legal-guidance/stalking-and-harassment.

Craigen, D., Diakun-Thibault, N. and Purse, R., 2014. *Defining cybersecurity*. Technology Innovation Management Review, 4(10).

Critcher, C. (2003). *Moral panics and the media*. Maidenhead: Open University Press.

D'Andrea, E., Ducange, P., Bechini, A., Renda, A. and Marcelloni, F., 2019. Monitoring the public opinion about the vaccination topic from tweets analysis. *Expert Systems with Applications*, *116*, pp.209-226.

Deelman, E., Stodden, V., Taufer, M. and Welch, V., 2019, June. Initial Thoughts on Cybersecurity And Reproducibility. In Proceedings of the 2nd International Workshop on Practical Reproducible Evaluation of Computer Systems (pp. 13-15). ACM.

Deibert, R.J. and Crete-Nishihata, M., 2012. Global governance and the spread of cyberspace controls. *Global Governance: A Review of Multilateralism and International Organizations*, *18*(3), pp.339-361.

Devroye, L., 1981. On the almost everywhere convergence of nonparametric regression function estimates. *The Annals of Statistics*, pp.1310-1319.

Dhillon, G. and Smith, K.J., 2019. Defining objectives for preventing cyberstalking. Journal of Business Ethics, 157(1), pp.137-158.

Donalds, C. and Osei-Bryson, K.M., 2018, June. An Ontological Approach to Classifying Cybercrimes in an ICT4D Context. In *International Conference on Design Science Research in Information Systems and Technology* (pp. 253-267). Springer, Cham.

Donalds, C. and Osei-Bryson, K.M., 2019. Toward a cybercrime classification ontology: A knowledge-based approach. *Computers in Human Behavior*, *92*, pp.403-418.

Donalds, C. and Osei-Bryson, K.M., 2020. Cybersecurity compliance behavior: Exploring the influences of individual decision style and other antecedents. *International Journal of Information Management*, *51*, p.102056.

Dowty, D.R., Wall, R. and Peters, S., 2012. *Introduction to Montague semantics* (Vol. 11). Springer Science & Business Media.

Dua, S. and Du, X., 2016. *Data mining and machine learning in cybersecurity*. Auerbach Publications.

Dueñas-Osorio, L., Craig, J.I. and Goodno, B.J., 2007. Seismic response of critical interdependent networks. *Earthquake engineering & structural dynamics*, *36*(2), pp.285-306.

Dueñas-Osorio, L., Craig, J.I., Goodno, B.J. and Bostrom, A., 2007. Interdependent response of networked systems. *Journal of Infrastructure Systems*, *13*(3), pp.185-194.

Duggan, M., “Online Harassment 2017,” Pew Research Center, 11 July 2017. [Online]. Available: http://www.pewInternet.org/2017/07/11/online-harassment-2017/. [Accessed 16 July 2019].

Edhlund, B. and McDougall, A., 2019. NVivo 12 Essentials. Lulu. com.

Efron,B. (1979) Bootstrap methods: another look at the jackknife. Ann. Stat., 7, 1–26.

Efron,B. et al. (1996) Bootstrap confidence levels for phylogenetic trees. Proc. Natl Acad. Sci., USA, 93, 13429–13434.

Eldén, L., 2007. *Matrix methods in data mining and pattern recognition*. Society for Industrial and Applied Mathematics.

Ellis, R. and Mohan, V. eds., 2019. *Rewired: Cybersecurity Governance*. John Wiley & Sons.

Ellis, C.A., McClelland, A.C., Mohan, S., Kuo, E., Kasner, S.E., Zhang, C., Khankhanian, P. and Balu, R., 2019. Cerebrospinal fluid in posterior reversible encephalopathy syndrome: implications of elevated protein and pleocytosis. *The Neurohospitalist*, *9*(2), pp.58-64.

Ellison, N.B., Steinfield, C. and Lampe, C., 2007. *The benefits of Facebook "friends:" Social capital and college students' use of online social network sites*. Journal of computer-mediated communication, 12(4), pp.1143-1168. Vancouver

Eterovic-Soric, B., Choo, K.K.R., Ashman, H. and Mubarak, S., 2017. Stalking the stalkers–detecting and deterring stalking behaviours using technology: A review. Computers & security, 70, pp.278-289.

Fan, W. and Gordon, M.D., 2014. The power of social media analytics. *Commun. Acm*, *57*(6), pp.74-81.

FBI, "Internet Crime Report (2016)," Internet Crime Complaint Center, 2016.

FBI, "Internet Crime Report (2017)," Internet Crime Complaint Center, 2017.

Felsenstein,J. (1985) Confidence limits on phylogenies: an approach using the boot-

strap. Evolution, 39, 783–791.

Felsom, M. and Boba, R.L. eds., 2010. Crime and everyday life. Sage.

Ferraty, F. and Vieu, P., 2006. *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media.

Finn, J. (2004). A survey of online harassment at a university campus. Journal of Interpersonal Violence, 19 (4), 468-483.

Fisher, B., Cullen, F. T., & Turner, M. G. (1999). The extent and nature of the sexual victimization of college women: A national level analysis (p. 323). Washington, DC: National Institute of Justice.

Fisher, B. S., Cullen, F. T., & Turner, M. G. (2002). Being pursued: Stalking victimization in a national study of college women. Criminology & Public Policy, 1(2), 257-308.

Fox, K. A., Nobles, M. R., & Fisher, B. S. (2016). A multi-theoretical framework to assess gendered stalking victimization: The utility of self-control, social learning, and control balance theories. Justice Quarterly, 33(2), 319-347.

Frommholz, I., Al-Khateeb, H.M., Potthast, M., Ghasem, Z., Shukla, M. and Short, E., 2016. On textual analysis and machine learning for cyberstalking detection. Datenbank-Spektrum, 16(2), pp.127-135.

FTC, “FTC Releases Annual Summary of Consumer Complaints (2017),” FTC, 2017.

P. Garg, R. Rani, and S. Miglani, “Analysis and visualization of professionals Linkedin data,” in the proceedings of International Conference on Emerging Research in Computing, Information, Communication and Applications, Springer, 31 July - 01 August 2015, pp. 1–9.

Garber,M. et al. (2001) Diversity of gene expression in adenocarcinoma of the lung. [Erratum (2002) Proc. Natl Acad. Sci. USA, 99, 1098.] Proc. Natl Acad. Sci. USA, 98, 13784–13789.

G. Jagannathan, R.N. Wright, Privacy-preserving distributed k-means clus- tering over arbitrarily partitioned data, in: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD’05, 2005, pp. 593–599.

Gibson, J.R., 2019. The Intersection of Cyberstalking, Gender, and Capable Guardianship (Doctoral dissertation, Arkansas State University).

Gilbert, E. and Karahalios, K., 2009, April. Predicting tie strength with social media. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 211-220).

Golder, S.A. and Macy, M.W., 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, *333*(6051), pp.1878-1881.

Goldberger, J., Hinton, G.E., Roweis, S. and Salakhutdinov, R.R., 2004. Neighbourhood components analysis. *Advances in neural information processing systems*, *17*.

Gordon, S. and Ford, R., 2006. On the definition and classification of cybercrime. Journal in Computer Virology, 2(1), pp.13-20.

Gordon, L.A. and Loeb, M.P., 2006. *Managing cybersecurity resources: a cost-benefit analysis* (Vol. 1). New York: McGraw-Hill.

Goodno, N. H. (2007). *Cyberstalking, a new crime: Evaluating the effectiveness of current state and federal laws*. Missouri Law Review, 72 (1), 1–74.

Goyal, R., Chandra, P. and Singh, Y., 2014. Suitability of KNN regression in the development of interaction based software fault prediction models. *Ieri Procedia*, *6*, pp.15-21.

Grabosky, P., 2000, April. Computer crime: A criminological overview. In *tenth United Nations Congress on the Prevention of Crime and the Treatment of Offenders, Vienna. http://www. aic. gov. au/conferences/other/compcrime/index. html*.

Gull, R., Shoaib, U., Rasheed, S., Abid, W. and Zahoor, B., 2016. *Pre-processing of twitter's data for opinion mining in political context. Procedia Computer Science*, 96, pp.1560-1570. Vancouver.

Guo, G., Ping, X. and Chen, G., 2006, August. A Fast Document Classification algorithm based on improved KNN. In *First International Conference on Innovative Computing, Information and Control-Volume I (ICICIC'06)* (Vol. 3, pp. 186-189). IEEE.

Hanna, A., Wells, C., Maurer, P., Friedland, L., Shah, D., & Matthes, J. (2013). Partisan alignments and political polarization online: A computational approach to understanding the French and US presidential elections. Proceedings of the 2nd Workshop on Politics, Elections and Data, 15–22.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). *The rise of "big data" on cloud computing: Review and open research issues.* Information Systems,47, 98-115

Hassanat, A.B., 2018. Two-point-based binary search trees for accelerating big data classification using KNN. PloS one, 13(11), p.e0207772.

Hassanat, A., 2018. Norm-based binary search trees for speeding up KNN big data classification. *Computers*, *7*(4), p.54.

Hassanat, A.B., Abbadi, M.A., Altarawneh, G.A. and Alhasanat, A.A., 2014. Solving the problem of the K parameter in the KNN classifier using an ensemble learning approach. *arXiv preprint arXiv:1409.0919*.

Haeussler, M., Zweig, A.S., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Hinrichs, A.S., Gonzalez, J.N. and Gibson, D., 2019. The UCSC genome browser database: 2019 update. *Nucleic acids research*, *47*(D1), pp.D853-D858.

Hamerly, G. and Elkan, C., 2004. Learning the k in k-means. *Advances in neural information processing systems*, *16*, pp.281-288.

He, H., Maple, C., Watson, T., Tiwari, A., Mehnen, J., Jin, Y. and Gabrys, B., 2016. The security challenges in the IoT enabled cyber-physical systems and opportunities for evolutionary computing & other computational intelligence.

Henson, B., Reyns, B.W. and Fisher, B.S., 2011. Security in the 21st century: Examining the link between online social network activity, privacy, and interpersonal victimization. *Criminal Justice Review*, *36*(3), pp.253-268.

Hillman, H., Hooper, C. and Choo, K.K.R., 2014. Online child exploitation: Challenges and future research directions. Computer Law & Security Review, 30(6), pp.687-698.

Hiltz, S.R., Diaz, P., & Mark, G. (2011). Social media and collaborative systems for crisis management, ACM Transactions on Computer-Human Interaction, Vol. 18 Issue 4, December 2011, ACM New York, NY, USA, DOI: https://doi.org/10.1145/2063231.

Horsman, G. and Conniss, L.R., 2015. An investigation of anonymous and spoof SMS resources used for the purposes of cyberstalking. Digital Investigation, 13, pp.80-93.

Hu, C., Jain, G., Zhang, P., Schmidt, C., Gomadam, P. and Gorka, T., 2014. Data-driven method based on particle swarm optimization and k-nearest neighbor regression for estimating capacity of lithium-ion battery. *Applied Energy*, *129*, pp.49-55.

Imandoust, S.B. and Bolandraftar, M., 2013. Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. *International Journal of Engineering Research and Applications*, *3*(5), pp.605-610.

Internetlivestats.com, “Internet Users,” Internet Live Stats, 2019.

Ishai, Y., Kushilevitz, E., Ostrovsky, R. and Sahai, A., 2008, May. Cryptography with constant computational overhead. In Proceedings of the fortieth annual ACM symposium on Theory of computing (pp. 433-442). ACM.

Jackson, K. and Bazeley, P., 2019. Qualitative data analysis with Nvivo. SAGE Publications Limited.

Jain, A.K., Murty, M.N. and Flynn, P.J., 1999. Data clustering: a review. *ACM computing surveys (CSUR)*, *31*(3), pp.264-323.

Jamshidi, Y. and Kaburlasos, V.G., 2014. gsaINknn: A GSA optimized, lattice computing knn classifier. *Engineering Applications of Artificial Intelligence*, *35*, pp.277-285.

Jarvis, L. and Macdonald, S., 2015. What is cyberterrorism? Findings from a survey of researchers. *Terrorism and Political Violence*, *27*(4), pp.657-678.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. Proceedings of the 9th WebKDD and 1st SNA-KDD2007 Workshop on Web Mining and Social Network Analysis, 56–65

Jiang, L., Wang, S., Li, C. and Zhang, L., 2016. Structure extended multinomial naive Bayes. Information Sciences, 329, pp.346-356.

Jin, D., Hannon, C., Li, Z., Cortes, P., Ramaraju, S., Burgess, P., Buch, N. and Shahidehpour, M., 2016. Smart street lighting system: A platform for innovative smart city applications and a new frontier for cyber-security. *The Electricity Journal*, *29*(10), pp.28-35.

J. Vaidya, C. Clifton, Privacy-preserving k-means clustering over vertically partitioned data, in: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD’03, 2003, pp. 206–215.

Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R. and Wu, A.Y., 2002. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, *24*(7), pp.881-892.

Kaplan, A.M. and Haenlein, M., 2010. Users of the world, unite! *The challenges and opportunities of Social Media. Business horizons*, 53(1), pp.59-68. Vancouver.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. Business Horizons, 53(1): 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

Kaplan, A.M. and Haenlein, M., 2011. The early bird catches the news: Nine things you should know about micro-blogging. *Business horizons*, *54*(2), pp.105-113.

Karyofyllidis, K., 2019. Sentiment Analysis for Twitter Users of America.

Kassner, M.E., 2015. *Fundamentals of creep in metals and alloys*. Butterworth-Heinemann.

Katal, A., Wazid, M. and Goudar, R.H., 2013, August. Big data: issues, challenges, tools and good practices. In *2013 Sixth international conference on contemporary computing (IC3)* (pp. 404-409). IEEE.

Kerr,M.K. and Churchill,G.A. (2001) Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. Proc. Natl Acad. Sci. USA, 98, 8961–8965.

Klimburg, A. (Ed.), National cyber security framework manual, NATO CCD COE Publications (2012).

K. P. Sinaga and M. Yang, "Unsupervised K-Means Clustering Algorithm," in *IEEE Access*, vol. 8, pp. 80716-80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

Kwon, O., Lee, N. and Shin, B., 2014. Data quality management, data usage experience and acquisition intention of big data analytics. *International journal of information management*, *34*(3), pp.387-394.

Le Charlier, B., Musumbu, K. and Van Hentenryck, P., 1991, June. A generic abstract interpretation algorithm and its complexity analysis. In ICLP (pp. 64-78).

Leskovec, J., 2011, March. Social media analytics: tracking, modeling and predicting the flow of information through networks. In *Proceedings of the 20th international conference companion on World wide web* (pp. 277-278). ACM.

Li, X.F., Dong, H.L., Huang, X.Y., Qiu, Y.F., Wang, H.J., Deng, Y.Q., Zhang, N.N., Ye, Q., Zhao, H., Liu, Z.Y. and Fan, H., 2016. Characterization of a 2016 clinical isolate of Zika virus in non-human primates. *EBioMedicine*, *12*, pp.170-177.

Li, L., Che, Y., Zhang, H., Li, T. and Yang, M., 2011, August. KNN text categorization algorithm based on LSA reduce dimensionality. In *2011 6th IEEE Joint International Information Technology and Artificial Intelligence Conference*(Vol. 2, pp. 72-75). IEEE.

Liu, H., Zhang, S., Zhao, J., Zhao, X. and Mo, Y., 2010, November. A new classification algorithm using mutual nearest neighbors. In *2010 Ninth International Conference on Grid and Cloud Computing* (pp. 52-57). IEEE.

Lipshutz, M. and Taylor, S.L., 1997. Comprehensive document representation. *Mathematical and Computer Modelling*, *25*(4), pp.85-93.

Lloyd, S.P., 1982. Least squares quantization in PCM. IEEE Trans. Inf. Theory. https://doi.org/10.1109/TIT.1982.1056489.

Loayza, N. and Schmidt-Hebbel, K., 2002. Monetary policy functions and transmission mechanisms: an overview. *Serieson Central Banking, Analysis, and Economic Policies, no. 4*.

L.P. Sheridan and T. Grant
Psychology, Crime and Law, 13 (6) (2007), pp. 627-640, 10.1080/10683160701340528.

Lyndon, A., Bonds-Raacke, J. and Cratty, A.D., 2011. College students' Facebook stalking of ex-partners. *Cyberpsychology, Behavior, and Social Networking*, *14*(12), pp.711-716.

Lyu, K. and Kim, H., 2016. Sentiment analysis using word polarity of social media. *Wireless Personal Communications*, *89*(3), pp.941-958.

Mack, Y.P., 1981. Local properties of k-NN regression estimates. *SIAM Journal on Algebraic Discrete Methods*, *2*(3), pp.311-323.

J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, no. 14, Oakland, CA, USA., 21 June 1967, pp. 281–297

Manovich, L., 2011. Trending: The promises and the challenges of big social data. *Debates in the digital humanities*, *2*(1), pp.460-475.

Marcum, C.D., Higgins, G.E. and Ricketts, M.L., 2014. Juveniles and cyber stalking in the United States: *An analysis of theoretical predictors of patterns of online perpetration. International Journal of Cyber Criminology*, 8(1). Vancouver.

Marcum, C.D., Higgins, G.E. and Ricketts, M.L., 2014. Sexting behaviors among adolescents in rural North Carolina*: A theoretical examination of low self-control and deviant peer association. International Journal of Cyber Criminology*, 8(2), p.68.

Marcum, C.D. and Higgins, G.E., 2019. Cybercrime. In *Handbook on Crime and Deviance* (pp. 459-475). Springer, Cham.

Marcum, A., 2019. *Cellular phone with keystroke entry nullification concurrent with vehicular motion*. U.S. Patent 10,320,966.

Mariam, A.O., Mohamed, M.G. and Frederic, S., 2013. A Survey of Data Mining Techniques for Social Network Analysis. School of Computing Science and Digital Media, Robert Gordon University, pp.1-25.

Mattord, C.L., 2009. Lay writers and the politics of theology in medieval England from the twelfth to fifteenth centuries. Georgia State University.

McCallum, K., 2007. Indigenous violence as a 'mediated public crisis. In *ANZCA 2007: Communications, Civics, Industry: Refereed Proceedings of the Australian and New Zealand Communication Association Conference 2007* (pp. 1-15). Australian and New Zealand Communications Association (ANZCA).

M.C. Doganay, T.B. Pedersen, Y. Saygin, E. Savas, A. Levi, Distributed privacy preserving k-means clustering with additive secret sharing, in: Proceedings of the 2008 International Workshop on Privacy and Anonymity in Information Society, PAIS'08, 2008, pp. 3–11.

McFarlane, L., & Bocij, P. (2005). An exploration of predatory behavior in cyberspace: Towards a typology of cyberstalkers. First Monday, 8. Retrieved Feb 18, 2006, from http://firstmon day.org/issues/issues8_9/mcfarlane/index.html.

McQuade, S.C., 2006. *Understanding and managing cybercrime*. Boston: Pearson/Allyn and Bacon.

Miller, T., Birch, M., Mauthner, M. and Jessop, J. eds., 2012. Ethics in qualitative research. Sage.

Moldagulova, A. and Sulaiman, R.B., 2017, May. Using KNN algorithm for classification of textual documents. In *2017 8th International Conference on Information Technology (ICIT)*(pp. 665-671). IEEE.

Moore, A., "Cyberstalking and Women: Facts and Statistics," thoughtco.com, 2018.

Moore, D. and Rid, T., 2016. Cryptopolitik and the Darknet. *Survival*, *58*(1), pp.7-38.

Narasimhan, M., Jojic, N. and Bilmes, J.A., 2005. Q-clustering. *Advances in Neural Information Processing Systems*, *18*, pp.979-986.

A. Nextmedia, “Social networks overview: Current trends and research challenges,” European Commission Information Society and Media, 2010.

Ni, J., Qiao, F., Li, L. and Di Wu, Q., 2012, July. A memetic PSO based KNN regression method for cycle time prediction in a wafer fab. In *Proceedings of the 10th World Congress on Intelligent Control and Automation* (pp. 474-478). IEEE.

N. Thabet and T. R. Soomro, “Big Data Challenges,” Journal of Computer Engineering & Information Technology, vol. 4, no. 3, 2015. https://doi.org/10.4172/2324-9307.1000133.

Nobles, M.R., 2019. Environmental crime and contemporary criminology: Making a difference. *American Journal of Criminal Justice*, *44*(4), pp.656-669.

Nobles, M. R., Reyns, B. W., Fox, K. A., & Fisher, B. S. (2014). Protection against pursuit: A conceptual and empirical comparison of cyberstalking and stalking victimization among a national sample. Justice Quarterly, 31(6), 986-1014.

NW3C, “Criminal Use of Social Media (2013),” NW3C, 2013.

NW3C, “Cyberstalking (March 2015),” NW3C, 2015

NW3C, “Disaster Fraud 2017,” NW3C, 2017

Ogilvie, E., 2000, December. The internet and cyberstalking. In *Proceedings of Criminal Justice Responses Conference, Sydney* (pp. 1-7).

Olmstead, K. and Smith, A., 2017. Americans and cybersecurity. *Pew Research Center*, *26*.

Olson, P.D., Cribb, T.H., Tkach, V.V., Bray, R.A. and Littlewood, D.T.J., 2003. Phylogeny and classification of the Digenea (Platyhelminthes: Trematoda). *International journal for parasitology*, *33*(7), pp.733-755.

Panlogic. “Cyber Crime.” *National Crime Agency*, 26 Feb. 2021, www.nationalcrimeagency.gov.uk/what-we-do/crime-threats/cyber-crime.

Panlogic. “Cyber Crime.” *National Crime Agency*, 12 Apr. 2022, https://www.nationalcrimeagency.gov.uk/what-we-do/crime-threats/cyber-crime.

Pandey, A.C., Rajpoot, D.S. and Saraswat, M., 2017. Twitter sentiment analysis using hybrid cuckoo search method. *Information Processing & Management*, *53*(4), pp.764-779.

Parsons-Pollard, N. and Moriarty, L.J., 2009. Cyberstalking: Utilizing what we do know. *Victims and Offenders*, *4*(4), pp.435-441.

Pena, J.M., Lozano, J.A. and Larranaga, P., 1999. An empirical comparison of four initialization methods for the k-means algorithm. *Pattern recognition letters*, *20*(10), pp.1027-1040.

Pham, D.T., Dimov, S.S. and Nguyen, C.D., 2005. Selection of K in K-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, *219*(1), pp.103-119.

Pietkiewicz, M. and Treder, M., 2018. Cyberstalking in social media–Polish view. Journal of Modern Science, 3, p.38.

Pratama, Y.A. and Effendi, S., 2019.
SELECTION FEATURES TO IMPROVE THE ACCURACY OF K-NEAREST NEIGHBOR.

Pratama, B.Y. and Sarno, R., 2015, November. Personality classification based on Twitter text using Naive Bayes, KNN and SVM. In *2015 International Conference on Data and Software Engineering (ICoDSE)* (pp. 170-174). IEEE.

Ragan, S.L., Wittenberg-Lyles, E.M., Goldsmith, J. and Reilly, S.S., 2008. Communication as comfort: Multiple voices in palliative care. Routledge.

Ramos, A., 2015. Sensor data security level estimation scheme for wireless sensor networks. Sensors, 15(1), pp.2104-2136.

R.C. Shorey, T.L. Cornelius and C. Strauss
Journal of Family Violence, 30 (7) (2015), pp. 935-942, 10.1007/s10896-015-9717-7.

Rejito, J., Atthariq, A. and Abdullah, A.S., 2021, January. Application of text mining employing k-means algorithms for clustering tweets of Tokopedia. In *Journal of Physics: Conference Series* (Vol. 1722, No. 1, p. 012019). IOP Publishing.

Reyns, B.W., Henson, B., Fisher, B.S., Fox, K.A. and Nobles, M.R., 2016. A gendered lifestyle-routine activity approach to explaining stalking victimization in Canada. *Journal of Interpersonal Violence*, *31*(9), pp.1719-1743.

Reyns, B.W., Henson, B. and Fisher, B.S., 2012. Stalking in the twilight zone: Extent of cyberstalking victimization and offending among college students. *Deviant Behavior*, *33*(1), pp.1-25.

Reyns, B.W., Henson, B. and Fisher, B.S., 2011. Being pursued online: Applying cyberlifestyle–routine activities theory to cyberstalking victimization. *Criminal justice and behavior*, *38*(11), pp.1149-1169.

Reyns, B.W., Fisher, B.S. and Randa, R., 2018. Explaining cyberstalking victimization against college women using a multitheoretical approach: Self-control, opportunity, and control balance. *Crime & Delinquency*, *64*(13), pp.1742-1764.

Rhee, H.S., Kim, C. and Ryu, Y.U., 2009. Self-efficacy in information security: Its influence on end users' information security practice behavior. *Computers & security*, *28*(8), pp.816-826.

Rossini,A. et al. (2003) Simple parallel statistical computing in R. UW Biostatistics Working Paper Series. Paper 193, University of Washington, WA.

Sebastiani, F., 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), pp.1-47.

S. Inc., “Number of social media users worldwide from 2010 to 2021 (in billions),” Statista: The Statistics Portal, New York, 2019.

Saha, B., & Srivastava, D. (2014). Data quality: The other face of Big Data. In Proceedings – International Conference on Data Engineering (pp. 1294–1297). https://doi.org/ 10.1109/ICDE.2014.6816764.

Saini, I., Singh, D. and Khosla, A., 2013. QRS detection using K-Nearest Neighbor algorithm (KNN) and evaluation on standard ECG databases. *Journal of advanced research*, *4*(4), pp.331-344.

A. Sechelea, T. Do Huu, E. Zimos, and N. Deligiannis, “Twitter data clustering and visualization,” in Telecommunications (ICT), 2016 23rd International Conference on. IEEE, 16 May 2016, pp. 1–5.

Schober, M. F., Pasek, J., Guggenheim, L., Lampe, C., & Conrad, F. G. (2016). Social Media Analyses for Social Measurement. Public Opinion Quarterly, 80(1), 180–211. http://dx.doi.org/10.1093/poq/nfv048.

Scrucca, L. and Raftery, A.E., 2018. clustvarsel: a package implementing variable selection for Gaussian model-based clustering in R. *Journal of Statistical Software*, *84*.

Sen, G. and Grown, C., 2013. Development crises and alternative visions: Third world women's perspectives. Routledge.

Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). Big Data, Digital Media, and Computational Social Science: Possibilities and Perils. The Annals of the American Academy of Political and Social Science, 659(1), 6–13. http://dx.doi.org/10.1177/0002716215572084.

A. Sharma and R. Rani, “Community detection and analysis of Twitter social data,” in the proceedings of 1st International IEEE Conference ‘INBUSH ERA 2015’ Theme: Futuristic trends in computational analysis and knowledge management, 25-27 February, 2015.

Sharma, R., 2020. Cyber Crime Phishing and it's security measures. *National Journal of Cyber Security Law*, *3*(1).

Sheldon, P., Rauschnabel, P. and Honeycutt, J.M., 2019. *The dark side of social media: psychological, managerial, and societal perspectives*. Academic Press.

Shen, Y., Hock Chuan, C., & Cheng, S. H. (2016). The Medium Matters: Effects on What Consumers Talk about Regarding Movie Trailers. In Proceedings of the International Conference on Information Systems

Shen, H., Goodall, J.C. and Hill Gaston, J.S., 2009. Frequency and phenotype of peripheral blood Th17 cells in ankylosing spondylitis and rheumatoid arthritis. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, *60*(6), pp.1647-1656.

Sheridan, L.P., Blaauw, E. and Davies, G.M., 2003. Stalking: Knowns and unknowns. Trauma, Violence, & Abuse, 4(2), pp.148-162.

Sheridan, L.P. and Grant, T., 2007. Is cyberstalking different?. *Psychology, crime & law*, *13*(6), pp.627-640.

Shi, K., Li, L., Liu, H., He, J., Zhang, N. and Song, W., 2011, September. An improved KNN text classification algorithm based on density. In *2011 IEEE International Conference on Cloud Computing and Intelligence Systems* (pp. 113-117). IEEE.

Shimodaira,H. and Hasegawa,M. (2001) CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics, 17, 1246–1247.

Shimodaira,H. (2002) An approximately unbiased test of phylogenetic tree selection. Syst. Biol., 51, 492–508.

Shimodaira,H. (2004) Approximately unbiased tests of regions using multistep- multiscale bootstrap resampling. Ann. Stat., 32, 2616–2641.

Shorey, R.C., Cornelius, T.L. and Strauss, C., 2015. Stalking in college student dating relationships: A descriptive investigation. *Journal of family violence*, *30*(7), pp.935-942.

Singer, P.W. and Friedman, A., 2014. *Cybersecurity: What everyone needs to know*. OUP USA.

Siomos, K.E., Dafouli, E.D., Braimiotis, D.A., Mouzas, O.D. and Angelopoulos, N.V., 2008. Internet addiction among Greek adolescent students. *CyberPsychology & Behavior*, *11*(6), pp.653-657.

Smith, R.G. and Urbas, G., 2001. Controlling fraud on the Internet: A CAPA perspective. Confederation of Asian and Pacific Accountants.

Smoker, M. and March, E., 2017. Predicting perpetration of intimate partner cyberstalking: Gender and the Dark Tetrad. *Computers in Human Behavior*, *72*, pp.390-396.

Snyder, F., 2001. Sites of criminality and sites of governance. Social & Legal Studies, 10(2), pp.251-256.Vancouver.

Soomro, T.R. and Hussain, M., 2019. Social Media-Related Cybercrimes and Techniques for Their Prevention. Applied Computer Systems, 24(1), pp.9-17.

Soranzo, N., Spector, T.D., Mangino, M., Kühnel, B., Rendon, A., Teumer, A., Willenborg, C., Wright, B., Chen, L., Li, M. and Salo, P., 2009. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nature genetics*, *41*(11), pp.1182-1190.

Southworth, C., Finn, J., Dawson, S., Fraser, C. and Tucker, S., 2007. Intimate partner violence, technology, and stalking. *Violence against women*, *13*(8), pp.842-856.

Spitzberg, B.H. and Hoobler, G., 2002. Cyberstalking and the technologies of interpersonal terrorism. *New media & society*, *4*(1), pp.71-92.

Stahl, F., Gaber, M.M. and Adedoyin-Olowe, M., 2014. A survey of data mining techniques for social media analysis. *Journal of Data Mining & Digital Humanities*, *2014*.

Stamm, M.C., Wu, M. and Liu, K.R., 2013. Information forensics: An overview of the first decade. *IEEE access*, *1*, pp.167-200.

Stratton, J., 1997. Cyberspace and the Globalization of Culture. *Internet culture*, pp.253-275

Strawhun, J., Adams, N., & Huss, M. T. (2013). The assessment of cyberstalking: An expanded examination including social networking, attachment, jealousy, and anger in relation to violence and abuse.

Stieglitz, S. and Dang-Xuan, L., 2013. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of management information systems*, *29*(4), pp.217-248.

Steinmetz, K.F. and Nobles, M.R. eds., 2017. *Technocrime and criminological theory*. Routledge.

Steinmetz, K.F., 2018. Introduction: Technocrime at the Margins. *Journal of Qualitative Criminal Justice & Criminology*.

Stieglitz, S., Mirbabaie, M., Ross, B. and Neuberger, C., 2018. Social media analytics–Challenges in topic discovery, data collection, and data preparation. *International journal of information management*, *39*, pp.156-168.

Subramanian, A., Narayan, R., Corsello, S.M., Peck, D.D., Natoli, T.E., Lu, X., Gould, J., Davis, J.F., Tubelli, A.A., Asiedu, J.K. and Lahr, D.L., 2017. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, *171*(6), pp.1437-1452.

Susarla, A., Oh, J.-H., & Tan, Y. (2012). Social Networks and the Diffusion of User-Generated Content: Evidence from YouTube. Information Systems Research, 23(1), 23–41. http://dx.doi.org/10.1287/isre.1100.0339.

Surya, P.P., Seetha, L.V. and Subbulakshmi, B., 2019, June. Analysis of user emotions and opinion using Multinomial Naive Bayes Classifier. In 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 410-415). IEEE.

Suzuki, R. and Shimodaira, H., 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, *22*(12), pp.1540-1542.

Suzuki,R. and Shimodaira,H. (2004) An application of multiscale bootstrap resampling to hierarchical clustering of microarray data: how accurate are these clusters? In proceedings by the Fifteenth International Conference on Genome Informatics (GIW 2004), p. P034.

T. R. Soomro and H. Wahba, “Perspectives of Cloud Computing: An Overview,” Proceedings 14th International Business Information Management Association (IBIMA) Conference on Global Business Transformation through Innovation and Knowledge Management, Istanbul, 2010.

Tang, J., Hu, X., Gao, H., & Liu, H. (2012). Unsupervised feature selection for linked social media data. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining − KDD ’12 (pp. 904–912). https://doi.org/10.1145/2339530.2339673.

Tao, D., Cheng, W.N. and Shang, W.Q., 2012, August. the research of knn text categorization algorithm based on eager learning [C]. In *2012 International Conference on industrial contral and electronics engineering*.

Teknomo, K., 2006. K-means clustering tutorial. *Medicine*, *100*(4), p.3.

Thelwall, M., 2017. Web indicators for research evaluation: A practical guide. San Rafael, CA: Morgan & Claypool. [An overview of all the steps needed from data collection to analysis and interpretation for web indicators, including practical advice.

Thomas, L., Buckland, S.T., Rexstad, E.A., Laake, J.L., Strindberg, S., Hedley, S.L., Bishop, J.R., Marques, T.A. and Burnham, K.P., 2010. Distance software: design and analysis of distance sampling surveys for estimating population size. *Journal of Applied Ecology*, *47*(1), pp.5-14.

Tjaden, P.G., 2014. Stalking and cyberstalking. *The encyclopedia of criminology and criminal justice*, pp.1-6.

Tokunaga, R.S., 2011. Social networking site or social surveillance site? Understanding the use of interpersonal electronic surveillance in romantic relationships. *Computers in human behavior*, *27*(2), pp.705-713.

Tokunaga, R.S. and Aune, K.S., 2017. Cyber-defense: A taxonomy of tactics for managing cyberstalking. *Journal of interpersonal violence*, *32*(10), pp.1451-1475.

Tran, L.Q., Moon, C.W., Le, D.X. and Thoma, G.R., 2001, July. Web page downloading and classification. In *Proceedings 14th IEEE Symposium on Computer-Based Medical Systems. CBMS 2001* (pp. 321-326). IEEE.

Uzialko, A. C., 2018. How Businesses Are Collecting Data (And What They're Doing With It), Business News Daily, August 3, 2018.

Vakilinia, I., Tosh, D.K. and Sengupta, S., 2017, July. Privacy-preserving cybersecurity information exchange mechanism. In *2017 International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS)* (pp. 1-7). IEEE.

Varmuza, K., Filzmoser, P., Hilchenbach, M., Krüger, H. and Silen, J., 2014. KNN classification—evaluated by repeated double cross validation: Recognition of minerals relevant for comet dust. *Chemometrics and Intelligent Laboratory Systems*, *138*, pp.64-71.

Van Laer, J. and Van Aelst, P., 2009. Cyber-protest and civil society: the Internet and action repertoires in social movements. *Handbook on internet crime*, *230254*.

Vayena, E., Brownsword, R., Edwards, S.J., Greshake, B., Kahn, J.P., Ladher, N., Montgomery, J., O'Connor, D., O'Neill, O., Richards, M.P. and Rid, A., 2016. Research led by participants: a new social contract for a new kind of research. *Journal of Medical Ethics*, *42*(4), pp.216-219.

Von Solms, R. and Van Niekerk, J., 2013. From information security to cyber security. *computers & security*, *38*, pp.97-102.

Walls, A., Perkins, E. and Weiss, J., 2013. Definition: Cybersecurity. *Retrieved from Gartner. com website: https://www. gartner. com/doc/2510116/def inition-cybersecurity*.

Welton-Mitchell, C., Bujang, N.A., Hussin, H., Husein, S., Santoadi, F. and James, L.E., 2019. Intimate partner abuse among Rohingya in Malaysia: Assessing stressors, mental health, social norms and help-seeking to inform interventions. *Intervention*, *17*(2), p.187.

Weissbrodt, D., 2013. Cyber-conflict, Cyber-crime, and Cyber-espionage. Minn. J. Int'l L., 22, p.347.

Whitman, M. and Mattord, H.J., 2014. Information security governance for the non-security business executive.

Williams, M.L. and Levi, M., 2017. Cybercrime prevention. *Handbook of crime prevention and community safety*, *454*.

Wright, P.J. and Tokunaga, R.S., 2016. Men's objectifying media consumption, objectification of women, and attitudes supportive of violence against women. *Archives of Sexual Behavior*, *45*(4), pp.955-964.

Wu, H., Gordon, M.D. and Fan, W., 2010. Collective taxonomizing: A collaborative approach to organizing document repositories. *Decision Support Systems*, *50*(1), pp.292-303.

Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Philip, S.Y. and Zhou, Z.H., 2008. Top 10 algorithms in data mining. *Knowledge and information systems*, *14*(1), pp.1-37.

Wu, X., Zhu, X., Wu, G.Q. and Ding, W., 2013. Data mining with big data. *IEEE transactions on knowledge and data engineering*, *26*(1), pp.97-107.

Yan, Z. and Xu, C., 2010, July. Combining KNN algorithm and other classifiers. In *9th IEEE International Conference on Cognitive Informatics (ICCI'10)* (pp. 800-805). IEEE.

Yang, X.S. and He, X., 2013. Bat algorithm: literature review and applications. *International Journal of Bio-inspired computation*, *5*(3), pp.141-149.

Yar, M., 2005. The Novelty of 'Cybercrime' An Assessment in Light of Routine Activity Theory. European Journal of Criminology, 2(4), pp.407-427. Vancouver

Yar, M., 2012. Crime, media and the will-to-representation: Reconsidering relationships in the new media age. *Crime, media, culture*, *8*(3), pp.245-260.

Yar, M., 2014. *The cultural imaginary of the internet: Virtual utopias and dystopias*. Springer.

Yar, M. and Steinmetz, K.F., 2019. Cybercrime and society. SAGE Publications Limited.

Yoo, S., Song, J. and Jeong, O., 2018. Social media contents based sentiment analysis and prediction system. *Expert Systems with Applications*, *105*, pp.102-111.

Yaqoob, I., Hashem, I.A.T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N.B. and Vasilakos, A.V., 2016. Big data: From beginning to future. *International Journal of Information Management*, *36*(6), pp.1231-1247.

Zeng, D., Chen, H., Lusch, R. and Li, S.H., 2010. Social media analytics and intelligence. *IEEE Intelligent Systems*, *25*(6), pp.13-16.

Zhang, Liping & Tang, Shanyu & Cai, Zhihua. (2014). Robust and efficient password authenticated key agreement with user anonymity for session initiation protocol-based communications. Communications, IET. 8. 83-91. 10.1049/iet-com.2012.0783.

Zhang, S., Cheng, D., Deng, Z., Zong, M. and Deng, X., 2018. A novel kNN algorithm with data-driven k parameter computation. *Pattern Recognition Letters*, *109*, pp.44-54.

Zhang, S., Liu, L., Hu, Y., Lv, Z., Li, Q., Gong, W., Sha, H. and Wu, H., 2017. Derivation of human induced pluripotent stem cell (iPSC) line from a 79 year old sporadic male Parkinson's disease patient. *Stem cell research*, *19*, pp.43-45.

Zhang, S., Wu, X. and Zhu, M., 2010, July. Efficient missing data imputation for supervised learning. In *9th IEEE International Conference on Cognitive Informatics (ICCI'10)*(pp. 672-679). IEEE.

Zhang, S., 2010. KNN-CF Approach: Incorporating Certainty Factor to kNN Classification. *IEEE Intell. Informatics Bull.*, *11*(1), pp.24-33.

Zhang, S., Li, X., Zong, M., Zhu, X. and Wang, R., 2017. Efficient kNN classification with different numbers of nearest neighbors. *IEEE transactions on neural networks and learning systems*, *29*(5), pp.1774-1785.

Zhang, S., Li, X., Zong, M., Zhu, X. and Cheng, D., 2017. Learning k for knn classification. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *8*(3), pp.1-19.

Zhao, D., Hu, X., Xiong, S., Tian, J., Xiang, J., Zhou, J. and Li, H., 2021. K-means clustering and kNN classification based on negative databases. *Applied Soft Computing*, *110*, p.107732.

Zhou L, Wang L, Ge X, Shi Q. A clustering-Based KNN improved algorithm CLKNN for text classification. In 2010 2nd International Asia Conference on Informatics in Control, Automation and Robotics (CAR 2010) 2010 Mar 6 (Vol. 3, pp. 212-215). IEEE.