Hierarchical Semantic Representations of Online News Comments for Emotion Tagging using Multiple Information Sources

Chao Wang[1], Ying Zhang[1⋆], Wei Jie[2], Christian Sauer[2], and Xiaojie Yuan[1]

[1] College of Computer and Control Engineering, Nankai University, P.R.China,
{wangchao, zhangying, yuanxiaojie}@dbis.nankai.edu.cn
[2] School of Computing and Engineering, University of West London, UK,
{wei.jie, christian.sauer}@uwl.ac.uk

Abstract. With the development of online news services, users now can actively respond to online news by expressing subjective emotions, which can help us understand the predilections and opinions of an individual user, and help news publishers to provide more relevant services. Neural network methods have achieved promising results, but still have challenges in the field of emotion tagging. Firstly, these methods regard the whole document as a stream or bag of words and can't encode the intrinsic relations between sentences. So these methods cannot properly express the semantic meaning of the document in which sentences may have logical relations. Secondly, these methods only use semantics of the document itself, while ignoring the accompanying information sources, which can significantly influence the interpretation of the sentiment contained in documents. Therefore, this paper presents a hierarchical semantic representation model of news comments using multiple information sources, called Hierarchical Semantic Neural Network (HSNN). In particular, we begin with a novel neural network model to learn document representation in a bottom-up way, capturing not only the semantics within sentence but also semantics or logical relations between sentences. On top of this, we tackle the task of predicting emotions for online news comments by exploiting multiple information sources including the content of comments, the content of news articles, and the user-generated emotion votes. A series of experiments and tests on real-world datasets have demonstrated the effectiveness of our proposed approach.

Keywords: emotion tagging, hierarchical semantic representation, multiple information sources, neural network

1 Introduction

Due to the development of the internet, the past decades have witnessed an explosive growth in different types of web services such as blogs, forums, social networks and online news services. Among these various types of web services, online news has been an important type of information that attracts billions of

⋆ Corresponding author

users to read and actively respond by making comments. Users often express subjective emotions like sadness, happiness and surprise in their comments. Extracting these emotions contained in the comments can help us understand the preferences and perspectives of users, and help online news publishers to provide users with more personalized services. Therefore, an automatic emotion tagging method for online news comments is strongly desirable.

Emotion tagging is a fundamental problem in the research area of opinion mining and sentiment analysis, which has attracted much attention in information retrieval and natural language processing communities[14][17]. The emotion tagging problem can be formulated as a multi-classification problem, which calls for identifying multiple emotion categories (e.g., happiness, sadness and angry, etc.) from user-generated content including product reviews, posts on blogs or social networks, comments in forums or comments in online news services.

The dominating approaches usually utilize machine learning algorithms to build a classifier with hand-crafted features. Since the performances of traditional machine learners are heavily dependent on feature representations[5], deep learning methods become more and more popular recently due to the ability to learn discriminative features from data automatically.

Despite the achievement of neural network approaches, there still are some challenges. Firstly, how to encode the intrinsic relations between sentences in the semantic meaning of a document. This is important for emotion tagging because relations such as causality and contrast have great influence on determining the meaning of a document. However, existing studies usually fail to effectively capture the intrinsic relations, since sentences influence the semantic meaning equally whether they are before or after the adversatives. Secondly, these methods only use semantics of the document itself, while ignoring the accompanying information sources, which can have significant influence on interpreting the sentiment of the document. In the news comment scenario, the comments are users' response to the news articles, thus the emotions of the comments are influenced by the content of the news articles obviously. Moreover, many online news websites provide a emotion voting service through which users can share their emotions after reading news articles. These user-generated emotion votes can naturally provide guidance for assigning emotion tags to comments.

Therefore, this paper presents a hierarchical semantic representation model of news comments using multiple information sources, called Hierarchical Semantic Neural Network (HSNN). Firstly we bring in a novel neural network model to learn a hierarchical semantic representation of documents which encodes not only the semantics between words in a sentence but also the relations between sentences. Further, we combine the representations of multiple information sources including the comments, the news articles and the user-generated emotion votes together and introduce a novel classification method utilizing this hierarchical semantic representation in order to improve the result of emotion predicting and tagging. A series of experiments and tests on real-world datasets have demonstrated that our HSNN demonstrated good performances in emotion tagging compared with a selection of baseline models.

The remainder of this paper is organized as follows. Section 2 gives a brief overview of some state-of-the-art research on emotion tagging and makes discussions regarding the differences between our work and previous works. In Section 3, we present our proposed approach HSNN including hierarchical semantic representation model of document and semantic representations using multiple information sources with their classifiers utilized. Experiments are shown in Section 4. We end the paper with conclusions and an outlook on future work.

2 Related Work

Emotion tagging has become an important subtask of opinion mining and sentiment analysis [14], which aims at identifying the emotion tag of a document (e.g., review of products[24][25][26], news article[1][2][12][21], news comment[29][30]). For a general survey, please refer to [17]. This paper focuses on emotion tagging for comments of online news.

Many machine learning techniques have been applied on sentiment classification, such as unsupervised learning techniques (e.g., [26]), supervised learning techniques (e.g., [18]) and semi-supervised learning techniques (e.g., [22]). Many studies now focus on designing an effective feature schema. On this basis, relevant features can be extracted and classifiers like SVM could be used to classify each text into emotion categories. Other than these methods using only words to classify text, prior works[1][2][21] asserted it is arguable that emotions should be linked to specific topics instead of a single keyword, and proposed emotion-topic models by incorporating a intermediate layer of emotion into LDA. Moreover, in Li's method[12], documents are not treated equally and influence the prediction at different levels, in order to reduce the impact of noisy documents. The weakness of the aforementioned methods are obvious. They regarded the document as a bag of words, and didn't take semantics of the document into account, while the sentiment of the document have close ties to the semantic meaning. At the meantime, some other studies analyse the emotion present in documents by considering semantics. Zhang *et al.*[28] brought in a Conditional Random Fields based model which take the context into account to encode the reviews. and mined the sentiment polarity to the products. Tang *et al.*[24][25] constructed a neural network model, which modelled user-comment and product-comment consistencies and rated numeric scores to products accordingly. Inspired by word embedding, [15][16] presented a batch of methods by using both local and global semantics to improve the performance on sentiment analysis. Differing from the aforementioned approaches, this paper presents a neural network model capturing both the semantics within sentence and relations between sentences to learn hierarchical document representation. Thus we can make full use of the semantic information to predict the sentiment of the documents.

On another hand, these methods only use the information of the document itself, while ignoring the accompanying information sources, which can significantly influence the interpretation of the sentiment contained in documents. This paper uses heterogeneous information sources to analyse the sentiment. To the best of our knowledge, the only work on emotion tagging for news com-

ments is Zhang's prior works[29][30], which used a fixed combination strategy to merge heterogeneous information sources, and employed traditional machine learning method to tag emotions for the comments of news. Our work differs from Zhang's work since we build our model based on artificial neural networks, instead of traditional machine learning. In addition, Zhang only uses two kinds of information sources, while we use more.

3 Hierarchical Semantic Neural Network

We now state the emotion tagging problem as follows: Given a set of users' comments on news along with the news articles and user-generated emotion votes of the news, we should identify the emotion tags of individual comments.

Furthermore, we formulate the problem setting as follows: Given a collection of comments C and a collection of news articles D, each $c \in C$ has its $d \in D$ which means c is made by a user after reading d. We also have a predefined emotion set $E = \{e_1, e_2, e_3, \cdots, e_K\}$ from which we assign emotion tag for each comment. Afterwards each news article d is accompanied by user-generated emotion votes $M_d = \{\mu_1, \mu_2, \mu_3, \cdots, \mu_K\}$ where $\mu_k \in \mathbb{R}$ is the count of votes over emotion e_k. On the top of this, we cast the emotion tagging problem into a multi-class classification problem that we classify a comment c into one emotion tag e_k of the emotion set, according to the content of the comment itself, the content of its news article d and the emotion votes M_d of d.

The problem involves three issues. First, we develop a hierarchical semantics representations model of the document, according to not only the contextual relations within sentence, but also the intrinsic relations between sentences to encode the semantic meaning of document. Second, we reconstruct comments as a combination of the semantic representations of the comment itself, its news article and the user-generated emotion vote of the news. Then we use this representation as feature to classify and assign emotion tags to comments.

3.1 Hierarchical Semantic Representation Model of the Document

We introduce our proposed hierarchical semantic representation model of the document in this section, which computes fixed length continuous vector representations for documents of variables length.

Words are the basic components of sentences, and sentences constitute documents structurally and semantically. The principle of compositionality[7] states that the meaning of a longer expression (e.g., a sentence or a document) comes from the meanings of its constituents and the rules used to combine them. Thus our method to compute the document representation can be divided into two steps. We first model sentence semantic representations by producing continuous sentence vectors from word vectors/representations. Then we use sentence semantic representations to get the final document semantic representations.

3.1.1 Sentence Semantic Representation

In order to model the sentence semantic representation, word embedding[3] is innovated to represent each word. According to word embedding, each word is

represented as a low dimensional, continuous and real-valued vector, all of which are stored in a matrix $L \in \mathbb{R}^{dim \times |V|}$, where dim is the dimension of word vectors and V is the vocabulary. The word embedding can be initialized randomly from a uniform distribution and learned as a parameter at the some time with the trainning of a neural network[10][23], or be pre-trained from text corpus with embedding learning algorithms[16][19][24]. We employ the latter method using *word2vec*[3] to make better use of semantic and grammatical associations of words.

After that, we apply a modified convolutional neural network (CNN) to compute representations of sentences. CNN are a state-of-the-art semantic model from sentiment classification and emotion tagging[10][11][24], and it can learn fixed length vectors for sentences of varying length, according to the words order in a sentence and doesn't depend on an external parse tree.

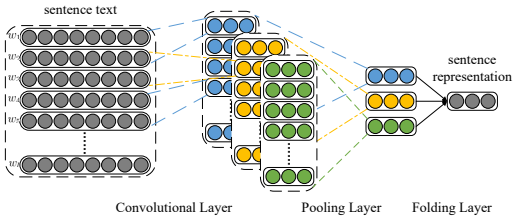

Fig. 1. Convolutional neural network for sentence semantic representation

Figure 1 shows the overview of our sentence method to capture the sentence semantic representation. The first lookup layer mapping words into low-dimensional vectors. The next layer performs convolutions over the embedded word vectors using filters with multiple sizes of windows. Next, we average-pool and average-fold the outputs of the convolutional layer into the representation.

We use different convolutional filters with different window widths to capture local semantics of various granularities to generate the sentence representation, which have been proven effective for sentiment classification and emotion tagging. For example, a convolutional filter with a window width of 3 essentially captures the semantics of a sentence in the perspective of trigram. In this paper, we use three different convolutional filters with widths of 3, 4 and 5 to encode the semantics of trigrams, 4-grams and 5-grams in a sentence.

Formally, given a sentence consisting of n words denoted as $\{w_1, w_2, w_3, \ldots, w_n\}$, l_{cf} is the window width of a convolutional filter cf, W_{cf} and b_{cf} is the shared parameters of linear layers of this filter. Each word w_i in the sentence is mapped to its word embedding $we_i \in \mathbb{R}^{dim}$ through word a embedding matrix $L \in \mathbb{R}^{dim \times |V|}$, where dim is the dimension of word embedding. The input of a linear layer is the concatenation of l_{cf} word embeddings in the window of this filter, which is denoted as $I_{cf} = \left[we_i; we_{i+1}; \ldots; we_{i+l_{cf}-1}\right] \in \mathbb{R}^{dim \cdot l_{cf}}$. The output of a linear layer is shown as follows:

$$O_{cf} = tanh\left(W_{cf} \cdot I_{cf} + b_{cf}\right), \tag{1}$$

[3] https://code.google.com/archive/p/word2vec/

where $W_{cf} \in \mathbb{R}^{l_{ocf} \times dim \cdot l_{cf}}$, $b_{cf} \in \mathbb{R}^{l_{ocf}}$, l_{ocf} is the length of the output of this convolutional layer, $tanh$ is the hyperbolic tangent to increase the non-linear property without affecting the receptive fields of the convolution.

Afterwards, we feed all the outputs of a convolutional filter into an average pooling layer to capture the overall semantics. Then we use an average fold layer to merge the outputs of different filters to get the final sentence representation.

3.1.2 Document Semantic Representation

Next, we introduce our method to generate a document representation from the obtained sentence vectors, utilizing long-short term memory model (LSTM).

Given a set of vectors of sentences, a simple and natural strategy to form a text vector is taking the average/max/min value of the sentence vectors as text vector. Obviously it can't capture complex relations such as causality and contrast between sentences since it totally ignores the order and logical relationship of sentences. Using convolutional neural network is an alternative to model local relations using its convolution with shared parameters partly. But this capability is considerably limited by the window size of the convolutional filter. The main idea behind recurrent neural network is to make use of sequential information of sentences. RNN is called recurrent because it performs the same task for every element of a sequence, with the output being depended on the previous computations. This helps it to encode the relations between sentences in long sequences, even if the two related sentences are far from each other in theory. Unfortunately, RNN suffers from gradient vanishing or exploding[4], which means gradients may grow or decay exponentially over long sequences. This makes it nearly impossible to model long-distance correlations in a sequence.

To solve this problem, we use a modified long-short term memory model. The transition function of LSTM used in this paper is shown as follows:

$$f_t = \delta\left(W_f \cdot [h_{t-1}; x_t] + b_f\right), \tag{2}$$

$$i_t = \delta\left(W_i \cdot [h_{t-1}; x_t] + b_i\right), \tag{3}$$

$$\tilde{C}_t = \delta\left(W_C \cdot [h_{t-1}; x_t] + b_C\right), \tag{4}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t, \tag{5}$$

$$h_t = \delta\left(W_h \cdot [h_{t-1}; x_t] + b_h\right) \odot tanh(C_t), \tag{6}$$

where x_t is the input vector of LSTM at the t-th step, in this section it's the t-th sentence semantic representation. f_t, i_t, W_f, W_i, b_f, b_i adaptively forget and update the information of hidden vector and input vector, W_C and b_C form the candidate vector, h_{t-1} is the hidden vector which represents the history status and maintains the accumulated knowledge of previous $t-1$ step, C_{t-1} and $\tilde{C}_t$ represent the old cell state and new candidate vector respectively at the t-th step, W_h and d_h help to update the hidden vector from the old hidden vector, input vector and cell state vector. As a side note, $\odot$ is element-wise multiplication of two vector, which means two vectors are multiplied element by element.

In classical LSTM[8], the last hidden vector is regarded as the text representation as shown in Figure 2. In this paper, we make a further extension called

Avg LSTM by using the average of all hidden vectors as text representation. Thus we can take considerations of the differences of semantics and sentiment relations between sentences and with different historical granularities.

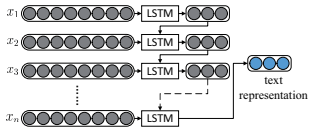

Fig. 2. Classical LSTM

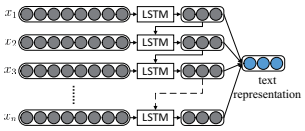

Fig. 3. Avg LSTM

3.2 Hierarchical Semantic Representation using Multiple Information Sources

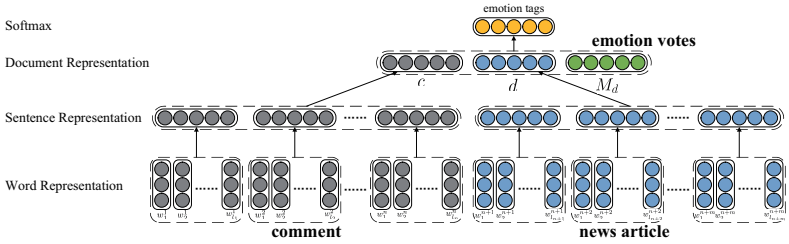

Fig. 4. The overview of our hierarchical semantic representation model using multiple information sources. w_j^i is the j-th word in the i-th sentence, l_i is the length of the i-th sentence, n and m are the numbers of sentences in the comment and the news article.

In this subsection, we introduce three heterogeneous information sources to mine sentiment of user comments, which are content of comments, content of news articles, and user-generated emotion votes of news articles.

The first and second information sources are hierarchical semantic representation of the contents of the comments and contents of the news articles. The comments are users' response to the news articles, thus the emotions of the comments are directly influenced by the content of the news articles. So we take contents of the news articles into consideration. For modelling the semantics of the comments and the news articles, we embed the content of comment c and news article d as continuous vectors $\bar{c} \in \mathbb{R}^{dim_c}$ and $\bar{d} \in \mathbb{R}^{dim_d}$ using the hierarchical semantic representation model in Section 3.1, where dim_c and dim_d are dimensions of the comment vector and the news vector respectively.

The last information source is derived from the emotion votes of the news articles generated by users. When tagging emotion for each comment, we can follow the normalized user-generated emotion votes of the news article to which the comment belongs. How likely the comment c of news article d will be tagged by emotion e_i according to the information of emotion votes can be denoted by $\bar{\mu_i}$. Hence we reconstitute the votes vector M_d through normalization as follows:

$$\bar{M_d} = \{\bar{\mu}_1, \bar{\mu}_2, \bar{\mu}_3, \cdots, \bar{\mu}_K\}, \bar{\mu}_i = \frac{\mu_i}{\sum_{j=1}^{K} \mu_j}. \tag{7}$$

On this basis, we define the final semantic representation vector of comment c as $rep(c) = \left[\bar{c}; \bar{d}; \bar{M}_d\right]$ and feed it to the classifier.

At last, an overview of our proposed hierarchical semantic representation model using multiple information sources is shown in Figure 4.

3.3 Sentiment Classification

In this section, we use hierarchical semantic representations using multiple information sources as discriminative features instead of handcrafted features which are used in traditional machine learning.

As shown in Section 3.2, the hierarchical semantic representation is the concatenation of semantic representation of comment, semantic representation of news article and continuous representation of user-generated emotion votes. On top of this, we introduce a *Softmax* classifier to transform the feature representations into conditional probabilities which can be interpreted as the probabilities of comments to be assigned into each emotion tag.

Given the i-th comment c_i in the corpus, the conditional probability that the comment should be associated with emotion $e_k (k = 1, 2, \cdots, K)$ within the set of emotion tags can be calculated as probability values with a *softmax* function.

$$P(e_k|c_i) = P(e_k|rep_i) = \frac{exp(\omega_k^T rep_i)}{\sum_{e_j \in E} exp(\omega_j^T rep_i)} \tag{8}$$

where c_i is the i-th comment, rep_i is the input hierarchical semantic representation feature of c_i, E is the set of emotion categories, ω is the matrix that transforms representation rep_i into a real-valued vector with dimension of $|E|$, ω_j is the combination parameter for each term with emotion e_j.

Afterwards, we train the model in a supervised way, where each comment in the training corpus is accompanied with its ground truth emotion tag. We introduce the cross-entropy error between ground truth sentiment distribution and predicted sentiment distribution as the objective loss function as follows:

$$J(\theta) = -\sum_{c \in C} \sum_{e \in E} P^g(e|c) \cdot log\left(P(e|c)\right), \tag{9}$$

where $c \in C$ is a comment, e is a emotion in the set of emotion categories E, $P(e|c)$ is the predicted distribution, $P^g(e|c)$ is the ground truth sentiment distribution with the same dimension of E, in which only the dimension corresponding to the ground truth is set to 1, and the others are set to 0.

We feed the cross-entropy error loss function into the back propagation algorithm to update the whole set of parameters of $\theta = [W_{cf}, b_{cf}, W_f, b_f, W_i, b_i, W_C, b_C, W_h, b_h, \omega]$ with stochastic gradient descent.

In this paper, we didn't enforce L2 norm constraints on parameters, instead we employ dropout[20][27] as a regularization method to reduce overfitting. The main idea of dropout is bringing in random removal of some units in a neural network during training, but keeping all of them during testing. Dropout involves a hyper parameter p, which means individual units are either "dropped out" of the network with the probability $1 - p$ or kept with the probability p in each

iteration, so that a reduced network is left to be trained in each iteration and the removed units keep their original weights.

Specifically, for the CNN part in this paper, before we feed I_{cf} into the convolutional layer, we add a dropout mask vector to the input vector to produce a dropout-modified input vector $\hat{I_{cf}}$ which is formulated as follows:

$$\hat{I_{cf}} = I_{cf} \odot m, \tag{10}$$

$$m_{(i)} \sim Bernoulli(p), \tag{11}$$

where m is the dropout mask with the same dimension of I_{cf}, and $m_{(i)}$ is the i-th element of m. Note that, m keeps changing for every I_{cf}. For the LSTM part, the hidden vector is also converted into a dropout-modified form similarly.

4 Experiment

In this section, we first introduce our experimental settings including the datasets used, evaluation metrics and baseline algorithms, then we present the experimental results with analysis and discussion.

4.1 Dataset

We collected the most-viewed news articles with their comments and user emotion votes in 6 months of 2011 from the *Society* channel of Sina News[4] and the *Entertainment* channel of QQ News[5]. We only use these Chinese datasets since we have not found similar services in English yet, but the proposed model is language independent. We randomly sampled news articles with their top-20 popular comments[6] and user-generated emotion votes as our training and testing datasets, which are referred as the Sina dataset and the QQ dataset respectively in the following pages. There are 5,185 comments, 369 news articles and 83,634 emotion votes in the Sina dataset, and 5,414 comments, 372 news articles and 993,089 emotion votes in the QQ dataset. Each comment is accompanied by its corresponding news articles and emotion votes of the news articles.

For the purpose of performance evaluation, emotion labels in both datasets are manually annotated. In Sina News and QQ News, even though users can tag articles with built-in emotion categories, the tag-systems are independent from the commenting systems so a tag cannot be paired with a specific comment. Thus we cannot utilize users tags as labels, instead, we just borrow the built-in emotion categories as predefined emotion categories in the annotating task. Due to the substantial laboring efforts, each dataset is annotated by only three annotators. The detailed statistic of labelled comments on the 8 emotions in Sina and QQ dataset are shown in Table 1. To test the annotating quality, 100 comments are randomly sampled from each dataset and a reviewer (not the annotator) annotated them blindly from the original labels. The number of consistent labels are 91 for the Sina dataset and 94 for the QQ dataset.

[4] http://news.sina.com.cn/society/

[5] http://ent.qq.com/

[6] If the number of comments was under 20, then we took all of them.

Table 1. The statistics of labeled comments of datasets.

Sina dataset

Emotion	Number	Proportion
Touched	905	17.45%
Sympathetic	614	11.84%
Bored	336	6.48%
Angry	1,752	33.79%
Amused	408	7.87%
Sad	654	12.61%
Surprised	196	3.78%
Fervent	320	6.17%

QQ dataset

Emotion	Number	Proportion
Happy	1,619	29.90%
Touched	139	2.57%
Sympathetic	641	11.84%
Angry	1,639	30.27%
Amused	563	10.40%
Sad	355	6.56%
Surprised	85	1.57%
Anxious	373	6.89%

4.2 Evaluation Metrics

In this paper, we apply two measures to compare the performances:

1. **Mean Reciprocal Rank (*MRR*)** Given a comment $c \in C$ with its ground truth emotion tag $\hat{e}_c$ and the predicted emotion ranking list L_c of c, let $rank_{L_c}(\hat{e}_c)$ be the position of $\hat{e}_c$ in L_c, MRR can be denoted as follows:

$$MRR = \frac{1}{|C|} \sum_{c \in C} \frac{1}{rank_{L_c}(\hat{e}_c)}. \tag{12}$$

2. **Accuracy (*Accu@m*)** Given a comment $c \in C$ with its ground truth emotion tag $\hat{e}_c$ and the predicted emotion ranking list $L_c@m$ including top-m emotions in L_c, $accu_c@m$ can be defined as follows:

$$accu_c@m = \begin{cases} 1, \hat{e}_c \in L_c@m \\ 0, \hat{e}_c \notin L_c@m \end{cases}. \tag{13}$$

and $Accu@m$ for the entire dataset is $Accu@m = \sum_{c \in C} accu_c@m/|C|$.

4.3 Baseline Methods

We compared the proposed HSNN with the following methods for emotion tagging with 10-fold cross validation on the two datasets.

1. In SVM+n-grams, we used bag-of-n-grams of comments as features and trained SVM classifier with LIBLINEAR[6].
2. WE, namely Word-Emotion method[21], is a generative model based on emotional dictionaries. It first builds the word-level and topic-level emotion dictionaries, then uses them to predict the emotions of given comments.
3. In RPWM, or Reader Perspective Weighted Model[12], comments are not treated equally and influence the prediction at different levels.
4. Standard CNN[11] and LSTM[8] are also implemented as baseline methods which are state-of-the-art technologies for semantics and sentiment analysis. Note that we used three convolutional filters with widths of 3, 4 and 5 for standard CNN as the same as our proposed HSNN.
5. Content-based Model (CM)[29] builds a supervised fixed combination classification model and uses traditional machine learning methods to predict emotions for the comments.
6. Finally, HSNN is our proposed model.

4.4 Comparison to Baselines

Table 2. Performances of emotion tagging using single information source.

	Sina dataset				QQ dateset			
	MRR	Accu@1	Accu@2	Accu@3	MRR	Accu@1	Accu@2	Accu@3
SVM+unigrams	0.6298	0.4455	0.6308	0.7615	0.6153	0.4256	0.6130	0.7549
SVM+bigrams	0.5901	0.4057	0.5734	0.6990	0.6028	0.4081	0.6094	0.7331
SVM+trigrams	0.5497	0.3528	0.5237	0.6627	0.5477	0.3084	0.5913	0.7118
WE	0.5687	0.3650	0.5587	0.7052	0.5340	0.3365	0.5077	0.6395
RPWM	0.5347	0.3356	0.4973	0.6512	0.5438	0.3638	0.5156	0.6206
Standard CNN	0.6166	0.4225	0.6668	0.7642	0.6326	0.4400	0.6172	0.7797
Standard LSTM	0.6414	0.4384	0.6856	0.7909	0.6833	0.4455	0.6317	0.8082
CM	0.6577	0.4838	0.6716	0.7810	0.6558	0.4907	0.6535	0.7636
HSNN_{CC}	**0.6841**	**0.5293**	**0.7478**	**0.8232**	**0.7046**	**0.4967**	**0.7077**	**0.8525**

The first set of experiments in this section is conducted to evaluate the performance of our proposed HSNN in comparison to the baseline methods using only the content of comments. Experimental results are shown in Table 2.

We can see that the SVM classifiers are very strong, which are almost the strongest among all baselines even though they nearly don't catch any linguistic information when the value of n is small. But with the increase of n, the bag-of-n-grams features become more and more sparse especially the comments part, since there are too few words in the comments. For example, the feature dimensions of unigrams, bigrams and trigrams on QQ dataset are 12,574, 90,687 and 158,741. This is also the reason why the performance of SVM with trigrams is the worst among three SVMs. We try to reduce the dimensions of features by only picking up emotion terms, but the performance shows no noticeable improvement.

WE is effective since it uses emotional dictionaries to predict the emotions of given comments. However, it only models comments as bag of words and doesn't take the semantic information of comments into account. RPWM is an improvement of WE, since it 1) jointly models emotions and topics by LDA, 2) calculates emotional entropy as document weights to reduce the impact of the noisy comments on the prediction. However, the results show no obvious improvement, we assume this is due to the fact that there is no significant difference between comments in the datasets used in this paper.

CM utilizes emotions terms[7] in the comments as features, and feeds them into a L2 regularization model. Since CM only takes considerations of the terms which are more likely to convey the emotions, it has a obviously better performance than the aforementioned baselines. From the comparison between CM and WE/RPWM, we also can tell that discriminative models usually have better performances and accuracies than generative models, which is proved in [9][13].

Standard CNN and LSTM outperform the vast majority of baseline methods significantly since they model the local semantics within the comment, from which we can tell that compositionality is important to understand the semantics and sentiment. However, there is still some room for improvement as long as the complex semantics, like the relations between sentences, are not captured well.

[7] Emotion terms can be extracted by several lexical resources developed for these tasks, such as NTU Sentiment Dictionary and Hownet.

HSNN_{CC}, which is our proposed model with single information source of content of comments, has an outstanding performance over all baseline methods, since it models not only the semantics within each sentence with modified CNN but also the relations between sentences with Avg LSTM. This gives HSNN the capability to model the complex semantics in documents. In addition, comparing HSNN_{CC} with Standard CNN and LSTM, we can tell that the logical relations between sentences do help understanding the sentiment and semantics of the whole comment positively.

Statistical significance tests have been conducted. HSNN_{CC} outperforms other methods with a confidence level of 0.95 on all datasets.

4.5 Effect of Multiple Information Sources

Table 3. Performances of HSNN with different information sources.

	Sina dataset				QQ dateset			
	MRR	Accu@1	Accu@2	Accu@3	MRR	Accu@1	Accu@2	Accu@3
HSNN_{CC}	0.6841	0.5293	0.7478	0.8232	0.7046	0.4967	0.7077	0.8525
HSNN_{CN}	0.6013	0.3732	0.4766	0.6725	0.5997	0.3575	0.4538	0.6732
HSNN_{UEV}	0.6019	0.3969	0.5299	0.6836	0.5995	0.3596	0.5077	0.7089
HSNN_{CC+CN}	0.6831	0.5232	0.7499	0.8357	0.7017	0.4986	0.7218	0.8557
HSNN_{CC+UEV}	0.7290	0.5859	0.8049	0.8713	0.7537	0.5403	0.7252	0.8947
HSNN_{CN+UEV}	0.6791	0.5105	0.7277	0.8515	0.6823	0.4860	0.6916	0.8291
HSNN_{CC+CN+UEV}	**0.7505**	**0.5905**	**0.8066**	**0.8904**	**0.7639**	**0.5605**	**0.7443**	**0.9049**

The second set of experiments is conducted to 1) find out whether every information source would be helpful for emotion tagging for comments, and 2) evaluate the performance of HSNN with different information sources.

HSNN_{CC}, HSNN_{CN} and HSNN_{UEV} are our proposed models with single information source, either of the comments, the news article or the user-generated emotion votes. HSNN_{CC+CN}, HSNN_{CC+UEV} and HSNN_{CN+UEV} are models with two information sources. HSNN_{CC+CN+UEV} is our integrated proposed model with all three information sources Finally.

From Table 3, we can see that HSNN_{CC} achieves the best performance compared to the other two single information source HSNN, which indicates the comments is more reliable and effective to predict the emotion of comments, since the comment is our object for emotion tagging obviously. HSNN_{UEV} is the second best, which means that the user-generated emotion votes are more useful than the news articles. This may be because the user emotion votes convey users' sentiments after reading the news articles more directly.

It can also be seen that HSNNs with multiple information sources generally outperform HSNNs with single information source, which shows that combining different information sources is more effective than using only one specific source of information. Furthermore, every information source is more or less helpful to understand the semantics and sentiment of comments.

Finally, HSNN_{CC+CN+UEV} yields the best performances, which clearly demonstrates that utilizing all three information sources is more effective than using only one or two specific sources of information. We can tell that each source of information provides a different perspective on emotion tagging, and respectively helps the model to achieve better prediction accuracy.

Statistical significance tests have been conducted. HSNN_{CC+CN+UEV} outperforms all other methods with a confidence level of 0.95 on all datasets.

4.6 Effect of Dropout

The final set of experiments is conducted to explore the effect of dropout.

Since a common value of dropout rate is $p = 0.5$ in practice[20], we designate it as our baseline, and effects of different dropout rates are measured by changes in Acc@1 compared with 0.5. Experimental data show that changes in MRR and Accu@2,3 have the same trend and a similar curve as Accu@1, so we only show the changes in Acc@1, which are shown in Figure 5.

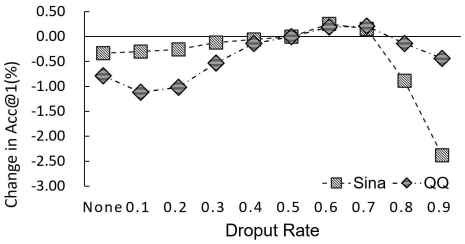

Fig. 5. Effect of dropout rate.

From Figure 5 we can see that non-zero dropout rates can improve the performance of emotion tagging at a range from 0.1 to 0.7, depending on datasets, which is consistent with the conclusions of previous research work[27]. In this paper, we choose $p = 0.6$ as our dropout rate during the experiment.

5 Conclusions and Future Work

In this paper, we proposed a novel methodology, namely Hierarchical Semantic Neural Network (HSNN), for emotion tagging for online news comments. Specifically, we developed a novel hierarchical semantic representation model to learn a semantic representation of a document based on both the semantics within a sentence and the relations between sentences. We also proposed a novel classification method utilizing the hierarchical semantic representation of multiple information sources. In this approach we use the information of not only the comment but also the accompanied news article and the user-generated votes to improve the classification accuracy of emotion tagging. The experimental results show that our approach outperforms the traditional approaches.

For possible future research, there are several assumptions, such as the improvement of HSNN to reorient the model to cross-domain and cross-language online news comments emotion tagging problem, or modelling the reading habits and emotional tendencies of individual users to improve the prediction accuracy.

6 Acknowledgement

This work is partially supported by National Natural Science Foundation of China under Grant No. 61402243 and National 863 Program of China under

Grant No. 2015AA015401. This work is also partially supported by Tianjin Municipal Science and Technology Commission under Grant No. 16JCQNJC00500 and No. 15JCTPJC62100.

References

1. Bao, S., Xu, S., Zhang, L., Yan, R., Su, Z., Han, D., Yu, Y.: Joint emotion-topic modeling for social affective text mining. In: 2009 Ninth IEEE International Conference on Data Mining. pp. 699–704. IEEE (2009)
2. Bao, S., Xu, S., Zhang, L., Yan, R., Su, Z., Han, D., Yu, Y.: Mining social emotions from affective text. IEEE Transactions on Knowledge and Data Engineering 24(9), 1658–1670 (2012)
3. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. Journal of Machine Learning Research 3(Feb), 1137–1155 (2003)
4. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks 5(2), 157–166 (1994)
5. Domingos, P.: A few useful things to know about machine learning. Communications of the ACM 55(10), 78–87 (2012)
6. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. Journal of Machine Learning Research 9(Aug), 1871–1874 (2008)
7. Frege, G.: Sense and reference. The Philosophical Review 57(3), 209–230 (1948)
8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation 9(8), 1735–1780 (1997)
9. Jordan, A.: On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. Advances in Neural Information Processing Systems 14, 841 (2002)
10. Kalchbrenner, N., Grefenstette, E., Blunsom, P.: A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188 (2014)
11. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
12. Li, X., Rao, Y., Chen, Y., Liu, X., Huang, H.: Social emotion classification via reader perspective weighted model. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence (2016)
13. Liang, P., Jordan, M.I.: An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In: Proceedings of the 25th International Conference on Machine Learning. pp. 584–591. ACM (2008)
14. Liu, B.: Opinion mining and sentiment analysis. In: Web Data Mining, pp. 459–526. Springer (2011)
15. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. vol. 1, pp. 142–150. Association for Computational Linguistics (2011)
16. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems. pp. 3111–3119 (2013)
17. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135 (2008)

18. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. vol. 10, pp. 79–86. Association for Computational Linguistics (2002)
19. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. vol. 14, pp. 1532–1543 (2014)
20. Pham, V., Bluche, T., Kermorvant, C., Louradour, J.: Dropout improves recurrent neural networks for handwriting recognition. In: Proceedings of the 14th International Conference on Frontiers in Handwriting Recognition. pp. 285–290. IEEE (2014)
21. Rao, Y., Lei, J., Wenyin, L., Li, Q., Chen, M.: Building emotional dictionary for sentiment analysis of online news. World Wide Web 17(4), 723–742 (2014)
22. Sindhwani, V., Melville, P.: Document-word co-regularization for semi-supervised sentiment analysis. In: 2008 Eighth IEEE International Conference on Data Mining. pp. 1025–1030. IEEE (2008)
23. Socher, R., Perelygin, A., Wu, J.Y., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1631–1642. Association for Computational Linguistics (October 2013)
24. Tang, D., Qin, B., Liu, T.: Document modeling with gated recurrent neural network for sentiment classification. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1422–1432 (2015)
25. Tang, D., Qin, B., Liu, T.: Learning semantic representations of users and products for document level sentiment classification. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. pp. 1014–1023 (2015)
26. Turney, P.D.: Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 417–424. Association for Computational Linguistics (2002)
27. Wu, H., Gu, X.: Towards dropout training for convolutional neural networks. Neural Networks 71, 1–10 (2015)
28. Zhang, K., Xie, Y., Yang, Y., Sun, A., Liu, H., Choudhary, A.: Incorporating conditional random fields and active learning to improve sentiment identification. Neural Networks 58, 60–67 (2014)
29. Zhang, Y., Fang, Y., Quan, X., Dai, L., Si, L., Yuan, X.: Emotion tagging for comments of online news by meta classification with heterogeneous information sources. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1059–1060. ACM (2012)
30. Zhang, Y., Zhang, N., Si, L., Lu, Y., Wang, Q., Yuan, X.: Cross-domain and cross-category emotion tagging for comments of online news. In: Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 627–636. ACM (2014)