Weighted Multi-label Classification Model for Sentiment Analysis of Online News

Xin Li*, Haoran Xie†, Yanghui Rao*‡, Yanjia Chen*,
Xuebo Liu*, Huan Huang* and Fu Lee Wang†
*Sun Yat-sen University, Guang Dong, China
Email: lixin77@mail2.sysu.edu.cn, raoyangh@mail.sysu.edu.cn, {chenyj79, liuxb5, huangh338}@mail2.sysu.edu.cn
†Caritas Institute of Higher Education, Tseung Kwan O, New Territories, Hong Kong SAR, China
Email: hrxie2@gmail.com, pwang@cihe.edu.hk
‡ The Corresponding Author

***Abstract*—With the extensive growth of social media services, many users express their feelings and opinions through news articles, blogs and tweets/microblogs. To discover the connections between emotions evoked in a user by varied-scale documents effectively, the paper is concerned with the problem of sentiment analysis over online news. Different from previous models which treat training documents uniformly, a weighted multi-label classification model (WMCM) is proposed by introducing the concept of "emotional concentration" to estimate the weight of training documents, in addition to tackle the issue of noisy samples for each emotion. The topic assignment is also used to distinguish different emotional senses of the same word at the semantic level. Experimental evaluations using short news headlines and long documents validate the effectiveness of the proposed WMCM for sentiment prediction.**

***Index Terms*—Sentiment analysis; Emotional concentration; Multi-label classification**

I. Introduction

The development of Web 2.0 technologies has been a great boon for the generation of online documents concerning user opinions. Sentiment analysis, also called opinion mining, is the field of studies that identify users' opinions, sentiments, appraisals, attitudes and emotions towards subjects [1]. The early studies of sentiment analysis [2] used supervised learning algorithms to classify polarity of reviews. The experimental results indicated that the algorithms performed worse on sentiment prediction than traditional text classification tasks. To further improve the performance, hybrid approaches which combined supervised, unsupervised and semi-supervised learning algorithms, have been developed to classify sentiments over multi-domain reviews [3].

Another stream of work focused on exploiting the sentiments of online news. For example, the task of "affective text analysis" in $SemEval$-2007 [4] was to annotate news headlines according to multiple emotion labels. As one of the top-performing lexicon-based algorithms on the above task, the SWAT algorithm [5] adopted a supervised approach to develop a word-emotion mapping dictionary. This dictionary was then used to determine the emotions of unlabeled news headlines. The emotion-term model [6] was also proposed to model word-emotion associations. These two models detected emotions at the word level. Due to that the same word in different subjects and contexts (or topics) may convey different attitudes [7], an emotion-topic model (ETM) was proposed to explore the particular emotions of topics [8]. Such a topic represents the real-world event, object, or abstract entity that indicates the subject or context of the sentiment [9]. ETM borrowed the machinery of latent topic models, such as the latent Dirichlet allocation (LDA) model [10], thus enabling different meanings of the same word to be distinguished. Recently, several multi-label topic models were proposed to detect emotions towards certain topics more accurately [11][12]. However, these models treated training documents uniformly and the documents that evoke prominent emotions in users are usually mixed with noisy documents that do not convey much affective meaning. Experimental results have shown that the performance of models without weighting for training documents is unstable, especially on the dataset with limited training instances or features [13].

In light of these considerations, we develop a weighted multi-label classification model (WMCM) for sentiment analysis of online news over varied-scale training documents. The main contributions of this work are as follows:

- Firstly, the developed model allows us to distinguish different emotional senses of the same word by introducing an additional topic layer.
- Secondly, we propose the concept of "*emotional concentration*" to estimate the weight of different training documents for each emotion.
- Finally, experimental evaluations using various baseline models and varied-scale datasets (i.e., an English corpus in $SemEval$-2007 tasks containing 246 news headlines in the training set only, and a large-scale Chinese news corpus containing 1,975,153 word tokens and 325,434 user ratings) validate the effectiveness of the proposed model for detecting emotions.

The remainder of this paper is organized as follows: We describe related work in Section II. We present the WMCM in Section III. Experimental evaluations are shown in Section IV. Finally, we present conclusions in Section V.

II. Related Work

In this section, we first review works related to sentiment analysis or opinion mining, before introducing domain-specific applications by adopting above sentimental and/or emotional techniques. We further summarize several works relating to multi-label classification, which will shed light on the background and the current state of research in this area.

A. Sentiment Analysis

Traditional sentiment analysis algorithms focused mainly on the polarity classification of reviews. Das and Chen [14] employed a manually created lexicon within a specific domain based on the previous study [15] to construct a general-purpose opinion lexicon that can be used across domains. However, the performance was dependent on certain key words. For instance, in the domain of stock market, the sentence "It's not a bear market" reflects a promising market but the negation word "not" may convey the reverse meaning, i.e., the model may predict this sentence as a negative sentence because of "not". Some paradigms of classification such as naïve Bayes, support vector machine, and maximum entropy were used to improve the performance of the task [2], but the results were not as good as those on text classification. Recently, the information of users was also used for sentiment analysis. Li et al. [16] incorporated the textual topic and user-word relationships into supervised topic modeling. Tan et al. [17] and Hu et al. [18] analyzed tweets by combining the textual information of them and user relationships in Twitter.

The research of sentiment analysis of online news originates from "affective text analysis" in $SemEval$-2007 tasks [4], in which, the dataset is a corpus of news headlines extracted from Google news and CNN. The aim of this task is to perform reader perspective emotion analysis in text data where one piece of text may evoke more than one kind of emotion. Preliminary works have focused mainly on exploiting the emotions of individual words. The SWAT system [5] employed the unigram model to annotate the emotional responses of news headlines, which scored the emotions of each word w as the average of emotions for every headline that contained w. The emotion-term (ET) model [6][8] is a variant of naïve Bayes, which directly models the word-emotion association by introducing generative model. Firstly, sample an emotion according to emotion frequency count. Then, sample a word for the given emotion under the priori probability $P(w|e)$. Finally, Bayesian method is used to estimate the posterior probability $P(e|d)$. The limitation of such models is that the same word may evoke positive attitude in one topic but negative in another.

Recently, the emotion topic model (ETM) [8] and several multi-label topic models [11][12] were developed to sentiment classification by introducing an additional topic layer between emotions and documents. However, due to the existence of noisy samples, the performance of existing models that weight training documents uniformly is quite unstable [13].

B. Domain-specific Applications

The above sentimental and emotional related techniques have been applied to various domains. An interesting and promising example is the stock price predictions: Li et al. [19] implemented a generic stock price prediction framework, and plugged in six sentimental models with different analyzing approaches to predict the stock prices in individual stock, sector and index levels. The predicting approach can be further enhanced and improved by employing sentimental analysis on a summarization model [20].

With the development of social media services, many approaches have been developed for sentiment analysis in social network environments [21]. For example, based on around 4 million tweets before and during the parliamentary election 2013 in Germany, Rill et al. [22] designed a system to detect emerging political topics in Twitter. It was observed that emerging topics can be extracted right after their occurrence in Twitter, in addition to be earlier than in Google Trends. Bell et al. [23] proposed an approach to social data analysis by exploring the usage of microblogging to manage interaction between humans and robots. The natural language processing techniques were employed in their approach to extract important features from text in social networks.

C. Multi-label Classification

The issue of supervised learning from multi-label data has attracted significant attention from researchers. For example, Ghamrawi and McCallum [24] explored multi-label conditional random field (CRF) classification models that directly parameterize label co-occurrences in multi-label classification to address the issue of the dependencies between labels. To tackle the problem that task is to output a label set whose size is unknown a priori for each unseen instance, a multi-label lazy learning approach named ML-kNN was presented, which is derived from the traditional k-nearest neighbor (kNN) algorithm [25]. Vens et al. [26] discussed several approaches to the induction of decision trees for Hierarchical multi-label classification (HMC), and further investigated their use in functional genomics through the experimental study. To assist Web-related tasks such as web page categorization or tag recommendation, Tang et al. [27] proposed the MetaLabeler to automatically determine the relevant set of labels for each instance without intensive human involvement or expensive cross-validation.

More recently, Read et al. [28] show that binary relevance-based methods have much to offer, and that high predictive performance can be obtained without impeding scalability to large datasets. Dembczynski et al. [29] elaborated more closely on the idea of exploiting label dependence, which contains two two types of label dependence (i.e., conditional and marginal dependence). Montanes et al. [30] proposed another technique of label dependencies called dependent binary relevance learning by combining properties of chaining and stacking. By combining multi-label classification models in the classifier chains family, Hong et al. [31] developed a

novel probabilistic ensemble framework for multi-label classification that is based on the mixtures-of-experts architecture. Comprehensive surveys on multi-label classification can be found in the recent literature [32][33].

III. Weighted Multi-label Classification Model

In this section, we detail our weighted multi-label classification model (WMCM) for sentiment analysis of online news. The notations of frequently-used terms are first defined. Then, we describe how to estimate the weight of documents and associate documents with words. Finally, we describe the method of predicting the sentiments of unlabeled documents.

A. Generic Framework

The objective of WMCM is to alleviate the issue of noisy training documents, and to distinguish different emotional senses of the same word. As illustrated in Figure 1, the general processes and components of WMCM follows.

- First, in the module of extracting topic information, the topics are distilled from the union set of training and testing documents by unsupervised topic models.
- Second, we estimate the emotional concentration of training documents by exploiting the distribution of emotions, so as to measure the contribution or weight of each document. This process is conducted in the modules of computing emotional concentration and weighting training documents.
- Third, the topic assignment is used as a "bridge" to associate documents with words at the semantic level. Then, the probability of each emotion conditioned to unlabeled documents is estimated by the Bayesian inference, and evaluated by the $F1$ measure.

B. Notation Definition

For convenience of describing our model, we define the following notations:

An online collection D consists of documents with word tokens from a vocabulary of W distinct items, and a set of ratings generated by online users over E kinds of emotion labels. For example, assume that the predefined five emotions are joy, anger, fear, sad and surprise (i.e., $E = 5$), a document d is voted on by 1 user over joy, 2 users over anger, 3 users over fear, 0 user over sad, and 4 users over surprise. Accordingly, the emotional responses of d can be denoted by $\{1, 2, 3, 0, 4\}$.

The whole corpus is modeled by K latent topics. The parameters θ_d and ψ_w are the multinomial topic distributions over document d and word w, respectively. The hyperparameter α determines the Dirichlet prior on θ_d and ψ_w, which can be interpreted as the number of unseen topic instances sampled from documents or words before training.

To enhance the topic learning on documents with limited features, an unordered word pair co-occurring in a fixed-size window of a word sequence (i.e., a "biterm") [34] is constructed for each document. Given a window size of generating biterms λ, the number of bitems is denoted as B. The number of biterms B increases as the size of the window λ grows. For instance, a document with three distinct words (w_1, w_2, w_3) will generate two biterms (w_1, w_2) and (w_2, w_3) when $\lambda = 2$. If λ is set to 3, three biterms (w_1, w_2), (w_2, w_3) and (w_1, w_3) will be generated. Table I summarizes the notations of frequently-used terms.

TABLE I: Notations of frequently-used terms.

Notation	Description
D	Collection of training documents
K	Number of topics
E	Number of emotion labels
W	Number of distinct word tokens
θ_d	Topic distributions of document d
ψ_w	Topic distributions of word w
α	Dirichlet prior of θ_d and ψ_b
B	Number of biterms
λ	Window size of generating biterms

C. Document Weight Estimation

Different from a typically single emotion expressed by an author, a distribution of user attitudes can be present across the span of a document [35]. For example, two training documents d_1 and d_2 may have the following number of user ratings over five emotions: {10, 0, 0, 0, 0} and {3, 2, 2, 2, 1} (E = 5). We first normalized user ratings and summed them to 1 for each document, i.e., the distribution of emotions conditioned to d_1 and d_2 are {1, 0, 0, 0, 0} and {0.3, 0.2, 0.2, 0.2, 0.1}. Although both d_1 and d_2 have the highest user ratings for the first emotion e_1, i.e., they belong to the same class according to majority vote, d_1 is more important than d_2 in terms of the degree of emotional discriminability. Thus, we further proposed the concept *emotional concentration* to estimate the weight of each document d in D_e, i.e., the collection of training documents that had the highest user ratings for emotion e. In the above example, we have $D_{e_1} = \{d_1, d_2\}$ and $D_{e_2} = D_{e_3} = D_{e_4} = D_{e_5} = \emptyset$.

Given E normalized user ratings for each document in non-decreasing order: $v_1 \leq v_2 ... \leq v_E$, we aim to capture the concentration of emotional distributions. Let $\bar{v}$ be the arithmetic mean of the above values, the minimum concentration corresponds to $v_1 = v_2 = ... = v_E = \bar{v}$, and $v_1 = v_2 = ... = v_{E-1} = 0, v_E = E\bar{v}$ corresponds to maximum concentration. Without loss of generality, the cumulative percentage of considered units up to the ith interval (i.e., F_i) and the cumulative percentage of the characteristic that belongs to the same first i units (i.e., Q_i) can be defined as [36]: $F_i = \frac{i}{E}$ and $Q_i = \frac{\Sigma_{j=1}^{i} v_j}{E\bar{v}}$, where $i = 1, ..., E$.

Accordingly, we estimate the weight / importance of each training document d in the collection D_e (i.e., the probability of d conditioned to the user's emotion of e) based on the differences $F_i - Q_i$ as follows:

$$P(d|e) = \frac{\Sigma_{i=1}^{E-1}(F_i - Q_i)}{\Sigma_{i=1}^{E-1} F_i}. \quad (1)$$

The above concentration index of emotional distributions satisfies the following properties:

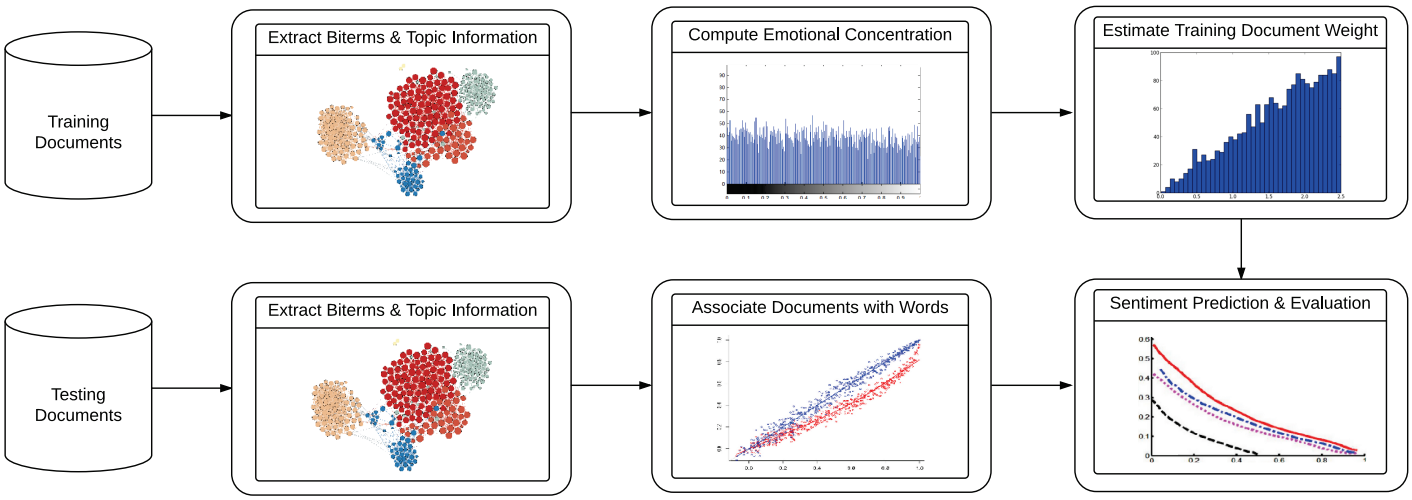


Fig. 1: The generic framework of weighted multi-label classification model (WMCM).

- The weight of document d equals 0 for minimum concentration when $F_i - Q_i = 0 (i = 1, ..., E)$.
- The weight of document d equals 1 for maximum concentration when $F_i - Q_i = F_i (i = 1, ..., E-1)$ and $F_E - Q_E = 0$.
- The weight increases as maximum concentration is approached and $0 < F_i - Q_i < F_i (i = 1, ..., E-1)$.

Compared to the index of entropy and many others that estimate the perplexity of a continuous variable only, the concept *emotional concentration* in our work applies to both continuous and ordinal variables [36]. Thus, we can also use the index of concentration to measure the document importance in the emotion ranking problem [35]. The values of document importance based on the index of concentration range from 0 to 1, with the highest being users voted for each emotion equally and the lowest being all users voted for a single emotion.

D. Associating Documents with Words

The aim of this part is to estimate the joint probability of training document d and word w. A straightforward method is based on the occurrence of words in documents; however, due to the fact of a single word may have emotional ambiguity, topic models are used as the "bridge" to associate documents with words accurately.

Many topic models such as latent Dirichlet allocation (LDA) [10] and probabilistic latent semantic analysis (pLSA) [37] have been used to extract the meaningful topics and alleviate the problem of ambiguity in sentiment analysis [38][39]. According to LDA, the word tokens are generated for each document d as follows:

- Choose θ_d from the Dirichlet distribution $Dir(\alpha)$;
- For each word token w:
  - Sample a topic z from the multinomial distribution $Multinomial(\theta_d)$;
  - Sample a word token w from the topic-word multinomial distribution.

Statistical techniques such as Gibbs sampling [40] can be used to topic learning. However, applying these models on short documents (e.g., news headlines) may suffer from the data sparsity problem [34]. The biterms were thus constructed to alleviate the above issue based on the aggregated word co-occurrence patterns in the corpus for discovering topics. For documents with limited features, the topic learning can be enhanced by generating biterms under a fixed-size window [34]. According to an approximate inference method, the topic distributions of document d can be estimated as follows:

$$\theta_d^{(z)} = \frac{n_d^{(z)} + \alpha}{B + K \times \alpha}, \tag{2}$$

where $n_d^{(z)}$ is the number of biterms, i.e., unordered word pairs in document d assigned to topic z.

After assigning topics to all biterms until convergence, we associate documents with words at the semantic level. The assumption is that words can hold semantic information as documents [41]. For instance, given a document d with 100 words, a word w with 100 instances in the corpus, and the number of topics is 4, we assume that there are 10 words in d assigned to topics 1, 2, 3, 4, respectively, and there are 10, 20, 30, 40 instances of w assigned to topics 1, 2, 3, and 4, respectively. Accordingly, the document d and word w both occurred most frequently in topic 4. If we instantiate the topic 4 with "machine learning", we can conclude that the main content of d is related to "machine learning", and in many cases w is also related to "machine learning". Inspired by this, we detect the relationships between documents and words at the topic level using the following two steps.

First, the above topic assignment is exploited to estimate the multinomial topic distributions over word w as follows:

$$\psi_w^{(z)} = \frac{n_w^{(z)} + \alpha}{\Sigma_{z'}(n_w^{(z')} + \alpha)}, \tag{3}$$

where $n_w^{(z)}$ is the number of instances of word w containing in all biterms assigned to topic z, and the consistent smoothing parameter α is the Dirichlet prior on θ_d and ψ_w.

Second, the cosine-similarity is employed to calculate the joint probability of d and w as follows:

$$P(d, w) = \frac{\theta_d \cdot \psi_w}{|\theta_d| \times |\psi_w|}. \tag{4}$$

Different from LDA, pLSA and other typical topic models that assign topics to bag of words, the aggregated patterns in the whole corpus are utilized by generating biterms. Previous works on topic discovery of short text have shown that topic models with biterms can learn more prominent and coherent topics than others [34]. This type of profile is useful for topic modeling over both short and long documents by tuning the window size λ.

E. Sentiment Prediction

Given an unlabeled document $\hat{d}$, the conditional probability of users' emotion e can be estimated by $P(e|\hat{d}) \propto P(e) \times P(\hat{d}|e)$. According to the maximum likelihood estimation,

$$P(e) = \frac{|D_e| + \beta}{|D| + E \times \beta}, \tag{5}$$

where $P(e)$ is the prior probability of emotion e, and β is a smoothing parameter used to avoid zero probability.

The estimation of $P(\hat{d}|e)$ is based on the conditional independence assumption that given users' emotion of e, each word w in $\hat{d}$ is generated independently. It is consistent to sentiment prediction of online news, in which the words of $\hat{d}$ are determined prior to the emotional responses triggered in users [41]. Thus, we have $P(\hat{d}|e) = \prod_{w \in \hat{d}} P(w|e)$. According to the Bayesian inference,

$$\begin{aligned} P(w|e) &= \Sigma_{d \in D_e} P(d|e) \times P(w|d) \\ &\propto \Sigma_{d \in D_e} P(d|e) \times P(d, w), \end{aligned} \tag{6}$$

where the probability of training document d conditioned to emotion e, i.e., $P(d|e)$ and the joint probability of d and w, i.e., $P(d, w)$ can be estimated by Eq. 1 and Eq. 4, respectively. This inference scheme weighted the training documents for each emotion in terms of the "emotional concentration", which effectively reduced the influence of noisy documents and enhanced the ability of classifiers to learn important features.

IV. Experiments

In this section, we detail the datasets, experiment design, and comparison with baselines.

A. Datasets

To evaluate the effectiveness and adaptiveness of the proposed model, we employ the following two datasets:

1) $SemEval$. An English dataset in $SemEval$-2007 tasks [4], which contains 1,250 news headlines extracted from Google news, CNN, and many others. In this dataset, each headline was manually scored in a fine-grained valence scale of 0 to 100 across 6 emotion labels (i.e., "anger", "disgust", "fear", "joy", "sad" and "surprise"). After pruning 4 items with the total scores equal to 0, we use the 246 headlines in the development set for training and the 1,000 in the testing set for evaluation.
2) $SinaNews$. A Chinese corpora consists of 4,570 news articles collected from the society channel of Sina [11]. The news headline, news body, and user ratings across 8 emotion labels (i.e., "touching", "empathy", "boredom", "anger", "amusement", "sadness", "surprise" and "warmness") were gathered. After pre-processing, there are 1,975,153 word tokens and 325,434 user ratings. Each document in the dataset has at least 6 word tokens and 1 user rating. Due to that adjacent news articles may have similar contexts, the 2,342 documents published from January to February, 2012 were used for training, and the 2,228 documents published from March to April, 2012 were used for testing.

The detailed information of the above datasets is shown in Table II, where the number of articles for each emotion label represents the amount of documents that had the highest ratings for that emotion.

TABLE II: Statistics of the datasets.

Dataset	Emotion label	# of articles	# of ratings
SemEval	anger	87	12042
	disgust	42	7,634
	fear	194	20,306
	joy	441	23,613
	sad	265	24,039
	surprise	217	21,495
SinaNews	touching	749	41,798
	empathy	225	23,230
	boredom	273	21,995
	anger	2,048	138,167
	amusement	715	43,712
	sadness	355	37,162
	surprise	167	11,386
	warmness	38	7,986

B. Experiment Design

In this part, we implemented the following baselines for comparison with our model WMCM:

1) SWAT system (SWAT): one of the top-performing systems on the $SemEval$ Affective Text Analysis task [42]. SWAT used the unigram model to annotate the emotional content of news headlines and scored the emotions of each word, which scored the emotions of each word w as the average of emotions of every headlines that w occurred [4][5].
2) Emotion term method (ET): a straightforward method to model the word-emotion associations [6]. ET follows the naïve Bayes (NB) method by assuming words are independently generated from emotional labels in two sampling steps. The difference between ET and NB is that emotion ratings is considered when calculating the $P(e)$ and $P(w|e)$.
3) Emotion topic model (ETM). The ETM model introduced an additional emotion layer into ET and LDA and

utilized the emotional distribution to reasonably guide the topic generation [8]. The parameters of ETM were set according to the description in [8].

4) Multi-labeled supervised topic model (MSTM) and Sentiment latent topic model (SLTM) [12]. MSTM began by generating topics from words, and then sampled emotions from each topic. SLTM, on the other hand, generated topics directly from user emotions.
5) Reader perspective weighted model (RPWM) [41]. The entropy and LDA were used to estimate the weight of documents and associate documents with words, respectively.

Table III presents the setting of parameters for our WMCM, where the values of α and other hyper-parameters on $SemEval$ (short documents) and $SinaNews$ (long documents) were specified by following [34]. The same holds for the number of iterations, which was set to 1,000. We fix the value of smoothing parameter β, since it has quite small impact on the prediction performance. The only parameter with different values for the two datasets was the size of window λ, which was determined by the averaged length of documents. Unless otherwise specified, all parameters of the baselines of ETM, MSTM and SLTM were set at default.

TABLE III: Parameters of WMCM.

Parameters	$SemEval$	$SinaNews$
α	$50/K$	$50/K$
β	0.01	0.01
λ	2	15

As mentioned earlier, the aim of sentiment analysis over online news is to mine emotions of users by predicting the probability of them conditioned on unlabeled news document, i.e., $P(e|\hat{d})$. A larger value of conditional probability means the document is more likely to arouse the corresponding emotion. To test the effectiveness of our model, we compare the predicted $P(e|\hat{d})$ with the actual distributions of emotions. In more detail, we can get a predicted label with the highest conditional probability and several top-ranked real labels, which are obtained by the ratings of real users. If the predicted label exists in the top-ranked real labels, this prediction is true, otherwise, the prediction is wrong. Assume that $Pred_{\hat{d}}$ is used to measure the quality of prediction, which is a binary variable and 0, 1 represent false and true respectively. It is calculated as follows:

$$Pred_{\hat{d}@n} = \begin{cases} 1, & e_p \in Top_{\hat{d}@n} \\ 0, & otherwise, \end{cases} \tag{7}$$

where e_p is predicated label and $Top_{\hat{d}@n}$ is the set of n top-ranked real labels derived from the emotional votes. The micro-averaged $F1$ measure was employed as the indicator of performance. The $F1$ measure equally weights precision and recall, and micro-averaging is one of the methods that can be used to compute a single aggregated measure when processing a collection with several two-class classifiers [43]. Micro-averaging pools per-document decisions across categories, and then computes an effectiveness measure on the pooled contingency table. Due to the very imbalanced distribution of documents in certain categories for both datasets (Table II), it is unnecessary to compute the $F1$ measure of each category or a macro-averaged $F1$ [43] that would take the average of $F1$ for all categories.

The computation of micro-averaged $F1$ in our work is based on $Pred_{\hat{d}@n}$, where n is set to 1, i.e., only the best match is the acceptable prediction. The equation is as follows:

$$\textit{Micro-averaged } F1 = \frac{\sum_{\hat{d} \in D_{test}} Pred_{\hat{d}@1}}{|D_{test}|}, \tag{8}$$

where D_{test} is the collection of testing documents, $Pred_{\hat{d}@1}$ is the prediction accuracy in the top-ranked emotional label as defined in Eq. 7. The larger value of Micro-averaged $F1$ indicates that the model is more effectiveness in terms of both precision and recall.

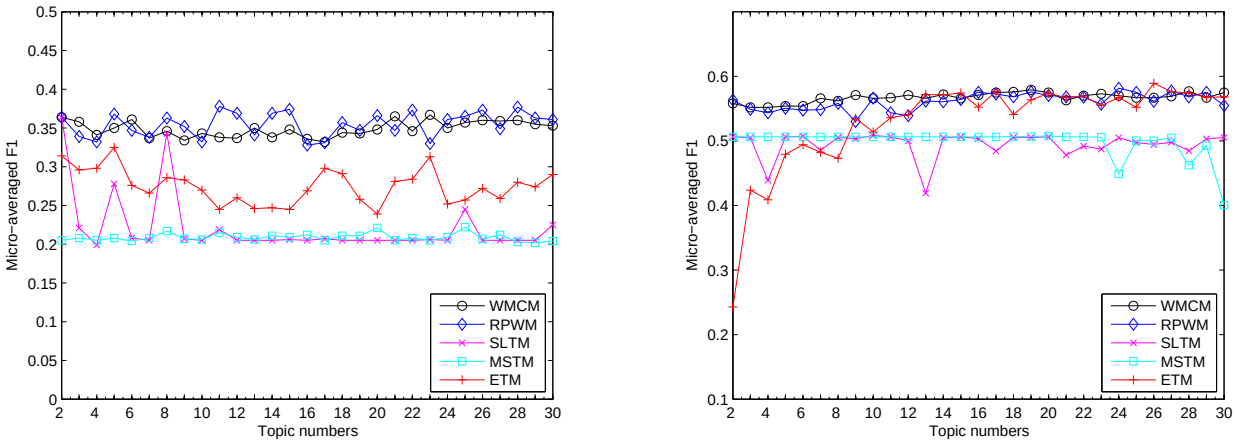
C. Comparison with Baselines

We compared the performance of our WMCM with SWAT and ET that do not take topic extraction into consideration, in addition to ETM, MSTM, SLTM and RPWM that exploit LDA in generating topics and predicting emotions. Since the number of topics may influence the performance of models involving topic learning, we varied the number of topics from 2 to 30 by following [6][8]. Figure 2 shows the performance of WMCM, RPWM, SLTM, MSTM and ETM when using those numbers of topics, from which we can observe that the proposed model WMCM achieved stable and better performance than the state-of-the-art baselines.

Compared to the baselines of SWAT, ET, ETM, MSTM and SLTM, the proposed WMCM improved 11.14%, 12.57%, 26.91%, 67.13%, 57.93% on $SemEval$, and 12.05%, 32.21%, 7.24%, 13.85%, 14.61% on $SinaNews$, respectively. In terms of the micro-averaged $F1$, the values of SWAT are 31.40% and 50.63% on $SemEval$ and $SinaNews$, and the values of ET are 31.00% and 42.91% on the two datasets. For other models exploiting topic learning, the mean value of the micro-averaged $F1$ over different numbers of topics was used. Although the baseline RPWM also performed well on average, it was less stable than our model in terms of the variance of $F1$ over those numbers of topics. The results validated the effectiveness of weighting training documents by exploiting the concept of emotional concentration, in addition to the association of documents with words at the semantic level.

To statistically evaluate the differences of performance, we performed two statistical tests on WMCM and each baseline model. The first one evaluated performance stability in terms of variances, and the second one compared performance in terms of means. We used the conventional significance level (i.e., p value) of 0.05.

First, we employed the analysis of variance (i.e., F-test) to evaluate the assumption of homoscedasticity (i.e., the homogeneity of variance). As SWAT and ET do not exploit latent topics, their performance is independent of the topic number. The F-test was thus conducted on WMCM and the baselines of RPWM, SLTM, MSTM and ETM (see Table IV). The



(a) SemEval (b) SinaNews

Fig. 2: Performance with different topic numbers.

results show that the differences in variances are statistically significant, with p values all less than 0.05. This suggests that the performance of WMCM is significantly more stable than that of RPWM, SLTM, MSTM and ETM when using a different number of topics.

TABLE IV: The p values of F-test on WMCM and baselines.

Models	$SemEval$	$SinaNews$
RPWM	0.0105	0.0023
SLTM	8.4E-11	5.2E-7
MSTM	0.0001	3.0E-8
ETM	3.2E-5	0.0000

Second, we conducted t-test to evaluate the assumption that the difference in performance between paired models had a mean value of zero (see Table V). The results indicate that the proposed WMCM outperformed the baselines of SLTM, MSTM, SWAT, ETM, ET and SWAT significantly, with p values much less than 0.05. Compared to the best-performing baseline of RPWM, although the performance of WMCM was slightly worse than that of RPWM on $SemEval$, the difference between them was not significant statistically (i.e., the p value equal to 0.0928). The performance of WMCM was better than that of RPWM on $SinaNews$, and the difference between them was statistically significant (i.e., the p value equal to 0.0363).

TABLE V: The p values of t-test on WMCM and baselines.

Models	$SemEval$	$SinaNews$
RPWM	0.0928	0.0363
SLTM	2.3E-17	2.7E-19
MSTM	1.8E-43	7.5E-17
ETM	8.0E-19	0.0080
ET	1.7E-18	2.8E-37
SWAT	2.9E-17	2.1E-27

V. Conclusion

Sentiment analysis is very useful for online service providers, which can help understand the preferences and perspectives of users and therefore facilitate the providers to provide users with more relevant and personalized services. Multi-label classification is one of the basic methods to associate documents with emotions. However, traditional classification algorithms [44] treat the samples / documents in the same class uniformly. Thus, most important samples for each emotion are usually mixed with noisy samples that do not convey much affective meaning. Our model mainly focused on taking the quality of each training sample into consideration when conducting sentiment analysis of online news.

In this work, we proposed the "emotion concentration" to alleviate the issue of noisy training documents, and exploited topic models to identify the emotional senses of the same word. Our model can be extended to other supervised learning algorithms such as support vector machines [45]. In the future, we will continue our work along the following directions:

1) We plan to evaluate the influence of the hyper-parameters on the performance of our model, in addition to develop an effective method of choosing the hyper-parameters automatically.
2) We estimated the weight of document by an index of "emotion concentration", which could be incorporated to a generalized supervised learning algorithm.
3) We plan to apply our weighted multi-label classification model to other fields such as stock prediction and movie or music recommendation.

Acknowledgment

The authors are thankful to the anonymous reviewers for their constructive comments and suggestions on an earlier version of this paper. The research described in this paper has been supported by the National Natural Science Foundation

of China (Grant No. 61502545), a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Grant No. UGC/FDS11/E06/14), and "the Fundamental Research Funds for the Central Universities" (Grant No. 46000-31610009).

REFERENCES

[1] Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers.

[2] Pang, B., Lee, L., and Vaithyanathan, S. (2002). *Thumbs up?: Sentiment classification using machine learning techniques.* In Proceedings of the International Conference on Empirical methods in Natural Language Processing (pp. 79-86). Association for Computational Linguistics.

[3] Kim, K., and Lee, J. (2014). *Sentiment visualization and classification via semi-supervised nonlinear dimensionality reduction*. Pattern Recognition, 47, 758-768.

[4] Strapparava, C., and Mihalcea, R. (2007). *Semeval-2007 task 14: Affective text*. In Proceedings of the 4th International Workshop on Semantic Evaluations (pp. 70-74). Association for Computational Linguistics.

[5] Katz, P., Singleton, M., and Wicentowski, R. (2007). *Swat-mp: the semeval-2007 systems for task 5 and task 14*. In Proceedings of the 4th International Workshop on Semantic Evaluations (pp. 308-313). Association for Computational Linguistics.

[6] Bao, S., Xu, S., Zhang, L., Yan, R., Su, Z., Han, D., and Yu, Y. (2009). *Joint emotion-topic modeling for social affective text mining*. In Proceedings of the 9th IEEE International Conference on Data Mining (pp. 699-704). IEEE.

[7] Quan, C., and Ren, F. (2010). *An exploration of features for recognizing word emotion*. In Proceedings of the 23rd International Conference on Computational Linguistics (pp. 922-930). Association for Computational Linguistics.

[8] Bao, S., Xu, S., Zhang, L., Yan, R., Su, Z., Han, D., and Yu, Y. (2012). *Mining social emotions from affective text*. IEEE Transactions on Knowledge and Data Engineering, 24(9), 1658-1670.

[9] Stoyanov, V., and Cardie, C. (2008). *Annotating Topics of Opinions*. In Proceedings of the 6th International Conference on Language Resources and Evaluation.

[10] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). *Latent dirichlet allocation*. Journal of Machine Learning Research, 3, 993-1022.

[11] Rao, Y., Li, Q., Wenyin, L., Wu, Q., and Quan, X. (2014). *Affective topic model for social emotion detection*. Neural Networks, 58, 29-37.

[12] Rao, Y., Li, Q., Mao, X., and Wenyin, L. (2014). *Sentiment topic models for social emotion mining*. Information Sciences, 266, 90-100.

[13] Rao, Y., Lei, J., Wenyin, L., Li, Q., and Chen, M. (2014). *Building emotional dictionary for sentiment analysis of online news*. World Wide Web, 17(4), 723-742.

[14] Das, S. R., and Chen, M. Y. (2007). *Yahoo! for Amazon: Sentiment extraction from small talk on the web*. Management Science, 53(9), 1375-1388.

[15] Yi, J., Nasukawa, T., Bunescu, R., and Niblack, W. (2003). *Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques*. In Proceedings of the 3rd IEEE International Conference on Data Mining (pp. 427-434). IEEE.

[16] Li, F., Wang, S., Liu, S., and Zhang, M. (2014). *Suit: A supervised user-item based topic model for sentiment analysis*. In Proceedings of the 28th AAAI Conference on Artificial Intelligence.

[17] Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., and Li, P. (2011). *User-level sentiment analysis incorporating social networks*. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1397-1405). ACM.

[18] Hu, X., Tang, L., Tang, J., and Liu, H. (2013). *Exploiting social relations for sentiment analysis in microblogging*. In Proceedings of the 6th ACM International Conference on Web Search and Data Mining (pp. 537-546). ACM.

[19] Li, X., Xie, H., Chen, L., Wang, J., and Deng, X. (2014). *News impact on stock price return via sentiment analysis*, Knowledge-Based Systems, 69: 14-23.

[20] Li, X., Xie, H., Song, Y., Zhu S., Li, Q., and Wang F. L. (2015). *Does summarization help stock prediction? A news impact analysis*, IEEE Intelligent Systems, 30(3): 26-34.

[21] Montejo-Raez, A., Diaz-Galiano, M., Martinez-Santiago, F., and Urena-Lopez, A. (2014). *Crowd explicit sentiment analysis*, Knowledge-Based Systems, 69: 134-139.

[22] Rill, S., Reinel, D., Scheidt, J., and Zicari, R. (2014). *Politwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis*, Knowledge-Based Systems, 69: 24-33.

[23] Bell, D., Koulouri, T., Lauria, S., Macredie, R., and Sutton, J. (2014). *Microblogging as a mechanism for human-robot interaction*, Knowledge-Based Systems, 69: 64-77.

[24] Ghamrawi, N., and Mccallum, A. (2005). *Collective multi-label classification*. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management (pp. 195-200). ACM.

[25] Zhang, M. L., and Zhou, Z. H. (2007). *ML-KNN: A lazy learning approach to multi-label learning*. Pattern Recognition, 40(7), 2038-2048.

[26] Vens, C., Struyf, J., Schietgat, L., Dzeroski, S., and Blockeel, H. (2008). *Decision trees for hierarchical multi-label classification*. Machine Learning, 73(2), 185-214.

[27] Tang, L., Rajan, S., and Narayanan, V. K. (2009). *Large scale multi-label classification via metalabeler*. In Proceedings of the 18th International Conference on World Wide Web (pp. 211-220). ACM.

[28] Read, J., Pfahringer, B., Holmes, G., and Frank, E. (2011). *Classifier chains for multi-label classification*. Machine Learning, 85(3), 333-359.

[29] Dembczynski, K., Waegeman, W., Cheng, W., and Hullermeier, E. (2012). *On label dependence and loss minimization in multi-label classification*. Machine Learning, 88(1-2), 5-45.

[30] Montanes, E., Senge, R., Barranquero, J., Quevedo, J. R., del Coz, J. J., and Hullermeier, E. (2014). *Dependent binary relevance models for multi-label classification*. Pattern Recognition, 47(3), 1494-1508.

[31] Hong, C., Batal, I., and Hauskrecht, M. (2015). *A generalized mixture framework for multi-label classification*. In Proceedings of the 2015 SIAM International Conference on Data Mining. SIAM.

[32] Gibaja, E., and Ventura, S. (2015). *A tutorial on multilabel learning*, ACM Computing Surveys, 47(3): Article 52.

[33] Zhang, M.-L., and Zhou, Z.-H. (2014). *A review on multi-label learning algorithms*, IEEE Transactions on Knowledge and Data Engineering, 26(8): 1819-1837.

[34] Cheng, X., Yan, X., Lan, Y., and Guo, J. (2014). *BTM: Topic modeling over short texts*. IEEE Transactions on Knowledge and Data Engineering, 26(12), 2928-2941.

[35] Lin, K. H. Y., and Chen, H. H. (2008). *Ranking reader emotions using pairwise loss minimization and emotional distribution regression*. In Proceedings of the International Conference on Empirical Methods in Natural Language Processing (pp. 136-144). Association for Computational Linguistics.

[36] Giudici, P. (2003). *Applied data mining: Statistical methods for business and industry*. In Wiley (pp. 39-41).

[37] Hofmann, T. (1999). *Probabilistic latent semantic analysis*. In Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (pp. 289-296). Morgan Kaufmann Publishers Inc..

[38] Lin, C., and He, Y. (2009). *Joint sentiment/topic model for sentiment analysis*. In Proceedings of the 18th ACM Conference on Information and Knowledge Management (pp. 375-384). ACM.

[39] Li, S., Huang, L., Wang, R., and Zhou, G. (2015). *Sentence-level emotion classification with label and context dependence*. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (pp. 1045-1053). Association for Computational Linguistics.

[40] Griffiths, T. L., and Steyvers, M. (2004). *Finding scientific topics*. Proceedings of the National Academy of Sciences, 101(suppl 1), 5228-5235.

[41] Li, X., Rao, Y., Chen, Y., Liu, X., and Huang, H. (2016). *Social emotion classification via reader perspective weighted model*. In Proceedings of the 30th AAAI Conference on Artificial Intelligence, poster paper.

[42] Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). *Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks*. In Proceedings of the International Conference on Empirical Methods in Natural Language Processing (pp. 254-263). Association for Computational Linguistics.

[43] Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1, p. 496). Cambridge University Press.

[44] Pang-Ning, T., Steinbach, M., and Kumar, V. (2006). *Introduction to data mining*. In Library of Congress (p. 74).

[45] Cortes, C., and Vapnik, V. (1995). *Support-vector networks*. Machine Learning, 20(3), 273-297.