International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 6 | ISSN : 2456-3307

Sentiment Analysis of Twitter Data Using Multi Class Semantic Approach

Dr. P. Sumathy[1], S. M. Muthukumari[2]
[1]Assistant Professor , BDU, Tiruchirappalli, Tamil Nadu, India
[2]M.Phil Scholar, BDU, Tiruchirappalli, Tamil Nadu, India

ABSTRACT

Growth in the area of opinion mining and sentiment analysis has been rapid and aims to explore the opinions or text present on different platforms of social media through machine-learning techniques with sentiment, subjectivity analysis or polarity calculations. Social media is a popular network through which user can share their reviews about various topics, news, products etc. People use internet to access or update reviews so it is necessary to express opinion. Sentiment analysis is to classify these reviews based on its opinion as either positive or negative category. First we have preprocessed the dataset to convert unstructured reviews into structured form. Then we have used lexicon based approach to convert structured review into numerical score value. In lexicon based approach we have preprocessed dataset using feature selection and semantic analysis. Stop word removal, stemming, and calculating sentiment score with help of Twitter dataset have been done in preprocessing part. Then we have applied classification algorithm to classify opinion as either positive or negative. Support vector machine algorithm is used to classify reviews where Multi class kernel SVM is modified by its hyper parameters and compared with existing Naivesbayes algorithm. So optimized SVM gives good result than SVM and naïve bayes. At last we have compared performance of all classifier with respect to accuracy.

Keywords : Social Media, Twitter, Machine learning techniques, Sentiment analysis, Multi class classification

I. INTRODUCTION

Sentiment Analysis is the process of finding the opinion of user about some topic or the text in consideration. It is also known as opinion mining. In other words, it determines whether a piece of writing is positive, negative or neutral. Now-a-days, people use microblogging sites to express their opinion about something. There are many popular microblogging sites like Facebook, Amazon etc. It has been useful in various domains like political, business and educational domain. Companies have been receiving polls about the products they manufacture. Previous research was to classify the sentiments into two classes i.e. positive and negative. But it was not useful for decision making. Here decision making refers to the solution for improving the positive opinion of the user regarding the domain in consideration. Hence need was to find the possible reasons behind sentiment variations to make decisions properly. There are many such examples in various domains like bollywood, political, healthcare and business domain. It seems very difficult to find the exact reasons behind sentiment variations as number of tweets are more than thousands for the target event.

There are several challenges in Sentiment Analysis. The first is an opinion word that is considered to be positive in one situation may be considered negative in another situation. A second challenge is that people don't always express opinions in a same way. Most traditional text processing relies on the fact that small differences between two pieces of text don't change the meaning very much. In Sentiment Analysis, however, "the picture was great" is very different from "the picture was not great". People can be contradictory in their statements. Most reviews

will have both positive and negative comments, which is somewhat manageable by analyzing sentences one at a time. However, in the more informal medium like twitter or blogs, the more likely people are to combine different opinions in the same sentence which is easy for a human to understand, but more difficult for a computer to parse. Sometimes even other people have difficulty understanding what someone thought based on a short piece of text because it lacks context. For example, "That movie was as good as its last movie" is entirely dependent on what the person expressing the opinion thought of the previous model. The user's hunger is on for and dependence upon online advice and recommendations the data reveals is merely one reason behind the emerge of interest in new systems that deal directly with opinions as a first-class object. Sentiment Analysis concentrates on attitudes, whereas traditional text mining focuses on the analysis of facts. There are few main fields of research predominate in Sentiment Analysis: Sentiment classification, feature based Sentiment classification and opinion summarization. Sentiment classification deals with classifying entire documents according to the opinions towards certain objects. Feature-based Sentiment classification on the other hand considers the opinions on features of certain objects. Opinion summarization task is different from traditional text summarization because only the features of the product are mined on which the customers have expressed their opinions.

II. DATASOURCE

People and companies across disciplines exploit the rich and unique source of data for varied purposes. The major criterion for the improvement of the quality services rendered and enhancement of deliverables are the user opinions. Blogs, review sites and micro blogs provide a good understanding of the reception level of products and services.

Blogs

The name associated to universe of all the blog sites is called blogosphere. People write about the topics they want to share with others on a blog. Blogging is a happening thing because of its ease and simplicity of creating blog posts, its free form and unedited nature. We find a large number of posts on virtually every topic of interest on blogosphere. Sources of opinion in many of the studies related to sentiment analysis, blogs are used.

Review Sites

Opinions are the decision makes for any user in making a purchase. The user generated reviews for products and services are largely available on internet. The sentiment classification uses reviewer's data collected from the websites like www.gsmarena.com (mobile reviews), www.amazon.com (product reviews), www. CNETdownload.com (product reviews), which hosts millions of product reviews by consumers

Micro-blogging

A very popular communication tool among Internet users is micro-blogging. Millions of messages appear daily in popular web-sites for micro-blogging such as Twitter, Tumbler, Face book. Twitter messages sometimes express opinions which are used as data source for classifying sentiment.

1.2 DATA CHARACTERISTICS

Twitter is a social networking and micro-blogging service that lets its users post real time messages, called tweets. Tweets have many unique characteristics, which implicates new challenges and shape up the means of carrying sentiment analysis on it as compared to other domains. Following are some key characteristics of tweets:

Message Length: The maximum length of a Twitter message is 140 characters. This is different from previous sentiment classification research that focused on classifying longer texts, such as product and movie reviews.

Writing technique: The occurrence of incorrect spellings and cyber slang in tweets is more often in comparison with other domains. As the messages are quick and short, people use acronyms, misspell, and

use emoticons and other characters that convey special meanings.

Availability: The amount of data available is immense. More people tweet in the public domain as compared to Facebook (as Facebook has many privacy settings) thus making data more readily available. The Twitter API facilitates collection of tweets for training.

Topics: Twitter users post messages about a range of topics unlike other sites which are designed for a specific topic. This differs from a large fraction of past research, which focused on specific domains such as movie reviews.

Real time: Blogs are updated at longer intervals of time as blogs characteristically are longer in nature and writing them takes time. Tweets on the other hand being limited to 140 letters and are updated very often. This gives a more real time feel and represents the first reactions to events. We now describe some basic terminology related to twitter:

Emoticons: These are pictorial representations of facial expressions using punctuation and letters. The purpose of emoticons is to express the user's mood.

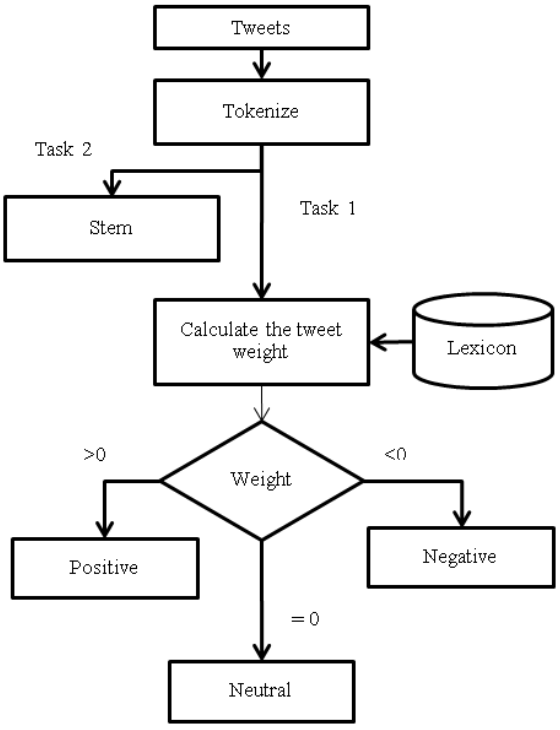


Fig 1: Basic Structure of Tweet analysis

Target: Twitter users make use of the "@" symbol to refer to other users on Twitter. Users are automatically alerted if they have been mentioned in this fashion.

Hash tags: Users use hash tags "#" to mark topics. It is used by Twitter users to make their tweets visible to a greater audience.

Special symbols: "RT" is used to indicate that it is a repeat of someone else's earlier tweet. The basic flow chart is shown in Figure 1.

III. RELATED WORK

A. L. Maas, et.al,….[1] present a model to capture both semantic and sentiment similarities among words. The semantic component of our model learns word vectors via an unsupervised probabilistic model of documents. However, in keeping with linguistic and cognitive research arguing that expressive content and descriptive semantic content are distinct, we find that this basic model misses crucial sentiment information.

B. Yang, et.al,…[2] analyzed the system for the ability to extract sentiment from text is crucial for many opinion-mining applications such as opinion summarization, opinion question answering and opinion retrieval. Accordingly, extracting sentiment at the fine-grained level (e.g. at the sentence- or phrase-level) has received increasing attention recently due to its challenging nature and its importance in supporting these opinion analysis tasks. In this paper, we focus on the task of sentence level sentiment classification in online reviews.

C. Lin, et.al,…[3] focused on document-level sentiment classification for general domains in conjunction with topic detection and topic sentiment analysis, based on the proposed weakly-supervised joint sentiment-topic (JST) model. This model extends the state-of-the art topic model latent Dirichlet allocation (LDA), by constructing an additional sentiment layer, assuming that topics are generated dependent on sentiment distributions and

words are generated conditioned on the sentiment-topic pairs.

Y. Jo, et.al,…[4] tackled two problems at once with a unified generative model of aspect and sentiment. Probabilistic topic models are suitable for the following two reasons: first, they provide an unsupervised way of discovering topics from documents (or aspects from reviews), and second, they result in language models that explain how much a word is related to each topic and possibly to a sentiment. The review is evaluating several aspects including the price, free upgrade, size, and sound, and each sentence expresses sentiment about one aspect. In the first sentence in the second paragraph, the words “monitor" and “bag" co-occur. In general, these two words are not closely related, but the co-occurrence of them signals that this sentence is evaluating the size of the monitor.

M. Dermouche, et.al,…[5] proposed model results in a 3-level output: topics, topic’s sentiments, and topic-sentiment evolution over time. It first serves as a traditional topic-discovering model able to extract the hidden topical structures from a document collection. Second, it models the association between topics and sentiments (the overall sentiment towards each of the extracted topics). Finally, it provides an efficient tool for tracking and visualizing the strength of topic-sentiment association over time.

D. Tang, et.al,…[6] propose a new model dubbed User Product Neural Network (UPNN) to capture user- and product-level information for sentiment classification of documents (e.g. reviews). Users and products are encoded in continuous vector spaces, the representations of which capture important global clues such as user preferences and product qualities.

X. Hu, et.al,…[7]extensively used to share information or opinions in various domains. With the growing availability of such an opinion-rich resource, it attracts much attention from those who seek to understand the opinions of individuals, or to gauge aggregated sentiment of mass populations. For example, advertisers may want to target users who are enthusiastic about a brand or a product in order to launch a successful social media campaign. Aid agencies from around the world would like to monitor sentiment evolutions before, during, and after crisis to assist recovery and provide disaster relief.

S. Liu, et.al,…[8] proposed the semi-supervised multiclass SVM model is formalized. Given a small amount of mixed labeled data from topics, it selects unlabeled tweets in the target topic, and minimizes the structural risk of labeled and selected data to adapt the sentiment classifier to the unlabeled data in a transductive manner. We set feature vector in the model into two parts: fixed common feature values and topic-adaptive feature variables. Topic-adaptive words as adaptive features are expanded and their values are updated in semi-supervised iterations to help transfer sentiment classifier.

Bayesian classification

The sentences that represent observations or attitude that is expressed as positive or negative are called as sentiments. The users post their tweets in twitter. These tweets are extracted in the form of unstructured data. The unstructured dataset is converted into structured form then extracts features from structured review. The features of the words are selected and then classification technique is applied on extracted features to classify them into its sentiment polarity that is namely either positive or negative. Feature words representation based on Naïve Bayes classifier is the main algorithm proposed by information retrieval researchers to represent text corpus.

Naïve Bayesian classifier

The Naive Bayesian classifier is based on Bayes’ theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods. Algorithm Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier assume that the effect of the value of

a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \text{ ---------- (1)}$$

$P(c|X) = P(x_1|c) * P(x_2|c)*....*P(x_n|c)*P(c)$
Where

- P(c|x) is the posterior probability of class (c, target)
  given predictor(x, attributes).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor

The Naivesbayesclassification only analyzed positive and negative reviews and provide reduced accuracy in classification.

IV. PROPOSED WORK

Sentiment Analysis (SA) and summarization has recently become the focus of many researchers, because analysis of online text is beneficial and demanded in many different applications. One such application is productbased sentiment summarization of multi-documents with the purpose of informing users about pros and cons of various products. This paper introduces a novel solution to target-oriented sentiment summarization and SA of short informal texts with a main focus on Twitter posts known as "tweets". We compare different algorithms and methods for SA polarity detection and sentiment summarization. We show that our hybrid polarity detection system not only outperforms the unigram state-of-the-art baseline, but also could be an advantage over other methods when used as a part of a sentiment summarization system

Support vector machine analyzes the data, define the decision boundaries and uses the kernels for computation which are performed in input space. The input data are two sets of vectors of size m each. Then every data represented as a vector is classified in a particular class. Now the task is to find a margin between two classes that is far from any document. The distance defines the margin of the classifier, maximizing the margin reduces indecisive decisions. SVM also supports classification and regression which are useful for statistical learning theory and it helps recognizing the factors precisely, that needs to be taken into account, to understand it successfully. It solves an optimization problem of finding the maximum margin hyperplane between the classes. This is basically required to avoid overfitting. Basically it is a linear classifier separating the classes which can be separated with the help of linear decision surfaces called hyperplanes. Foe classes having binary features SVM draws a line between the classes and for classes having multiple features hyperplanes are drawn. However it can be used for classifying the non linear data also by transforming the feature space into the higher dimensional space so that non linear data in higher dimensional can be separated easily by a hyperplane. This transformation is made easy with the help of Kernel-trick. With the help of kernels it is not necessary to calculate all the dimensions when transform and calculation of hyperplane can be done in the same lower dimensional feature space. Kernels are not used only for this purpose but also for making the calculation easier in case of many features. Various kernels are used by machine learning approaches e.g. RBF(Radial Basis Function), Linear kernel, Poly kernel etc.

In this paper, we proposed multi class SVM with 7 classes such as positive, negative, neutral, happy, hate, love and fun classes. We can construct feature vector list as seven and similar words. The pseudo code of the algorithm is shown as follows:

Input: Labeled Dataset
Output: positive, negative, neutral, happy, hate, love and fun polarity with synonym of words and similarity between words

Step1: Pre-Processing the tweets:
Pre-processing ()
Remove URL:
Remove special symbols
Convert to lower:
Step2: Get the Feature Vector List:
For w in words:
Replace two or more words

Strip:
If (w in stopwords)
Continue
Else: Append the file
Return feature vector

Step3: Extract Features from Feature Vector List:
For word in feature list
Features=word in tweets_words
Return features

Step4: Combine Pre-Processing Dataset and Feature Vector List
Pre-processed file=path name of the file
Stopwords=file path name
Feature Vector List=file path of feature vector list

Step5: Training the step 4 Apply classifiers classes
Step6: Find Synonym and Similarity of the Feature Vector

For every sentences in feature list (n=7)
Extract feature vector in the tweets ()
For each Feature Vector: x
For each Feature Vector: y
Find the similarity(x, y)
If (similarity>threshold)
Match found
Feature Vector: x= Feature Vector: y
Classify (x, y)

Print: sentiment polarity with similar feature words.

The above pseudo code is used to predict the polarity about multiclass based on features extract from twitter datasets. These features can be matched with vector to predict the similarity threshold and labeled as class name.

The proposed work is shown in Fig 3.

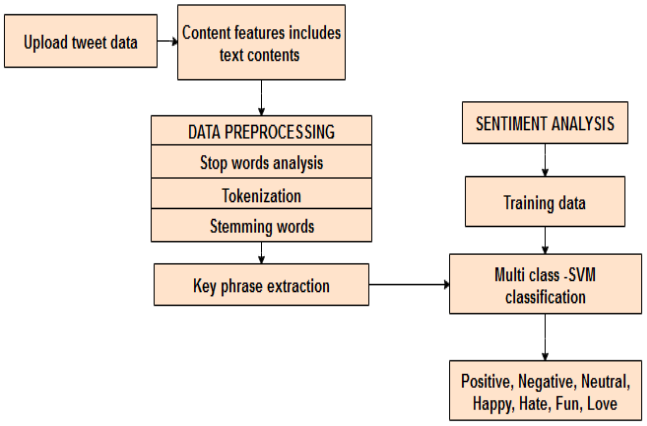


Fig 3: Proposed Framework

V. EXPERIMENTAL RESULTS

We used the twitter dataset in the form of CSV file and upload the datasets in R tool and analyze multi class SVM using R coding developed in R studio. The output results are shown in fig 4.

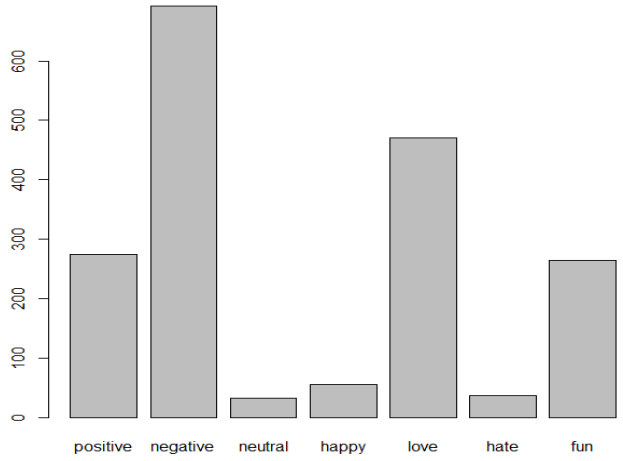


Fig 4: Pictorial Representation Of Twitter Data Set

The performance of the system is evaluated using Precision, Recall and F-measure.

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{---------- (2)}$$

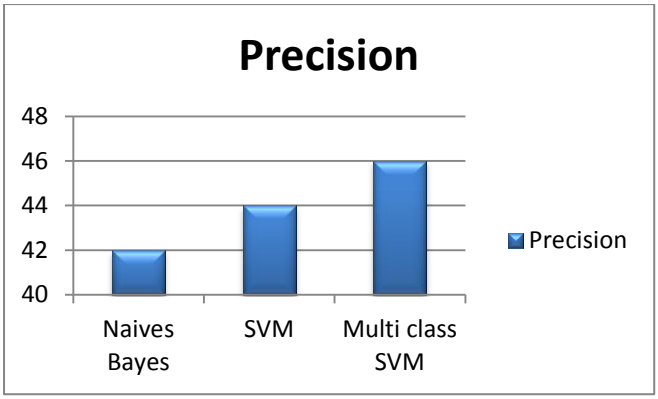
$$\text{Recall} = \frac{TP}{TP+FN} \quad \text{---------- (3)}$$

$$\text{F measure} = 2^* \frac{Precision*Recall}{Precision+Recall} \quad \text{-------- (4)}$$

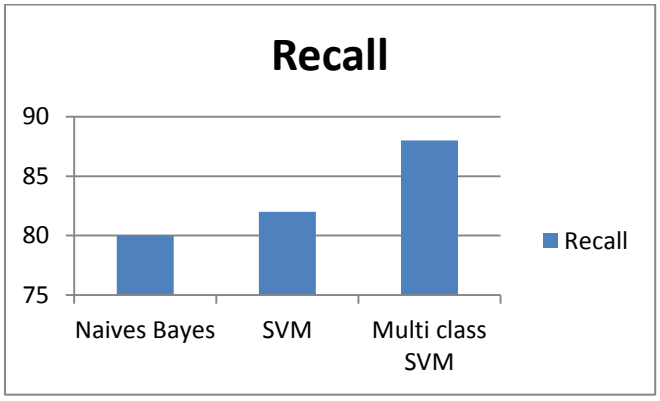
The performance evaluation result is shown in following table 1 and shows in fig 3.

Algorithm/ Performance measures	Precision	Recall	F- measure
Naives Bayes	42	80	55
SVM	44	82	57
Multi class SVM	46	88	60

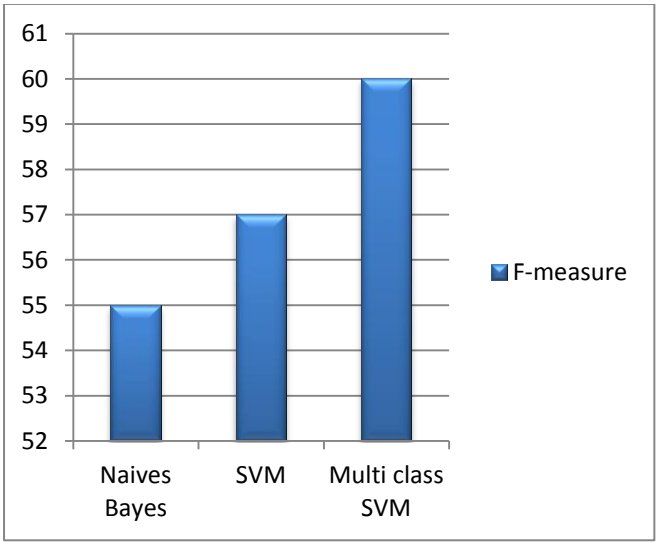
Table 1: Performance Table



(a)



(b)



(c)

Fig 5: (a) Precision (b) Recall (c) F-measure chart

From the above calculation, Multi class SVM provide high level F-measure values than the existing Naives Bayes and SVM algorithm

VI. CONCLUSION

In this paper, we conclude that to analyze machine learning techniques with semantic analysis for classify the reviews that are extracted from Twitter. The main objective is to handle large amount of reviews with multi class classification. The Multi class SVM algorithm can classify the various classes such as positive, negative, happy, love and fun. The proposed algorithm provides better results than the naivesbayes and SVM algorithm in terms of F-measure calculation. In future we can extend the approach to implement sentiment analysis with semantic approach in deep learning algorithms.

VII. REFERENCES

1. L Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in Proceedingsof the 49th Annual Meeting of the Association for ComputationalLinguistics: Human Language Technologies - Volume 1, ser. HLT'11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 142–150.
2. Yang and C. Cardie, "Context-aware learning for sentence-level sentiment analysis with posterior regularization," in Proceedingsof the 52nd Annual Meeting of the Association for ComputationalLinguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume1: Long Papers, 2014, pp. 325–335.
3. C Lin, Y. He, R. Everson, and S. Ruger, "Weakly supervised joint sentiment-topic detection from text," IEEE Transactions onKnowledge and Data Engineering, vol. 24, no. 6, pp. 1134–1145, Jun. 2012.
4. Y Jo and A. H. Oh, "Aspect and sentiment unification model for online review analysis," in Proceedings of the fourth ACM internationalconference on Web search and data mining, ser. WSDM'11. New York, NY, USA: ACM, 2011, pp. 815–824.
5. M Dermouche, J. Velcin, L. Khouas, and S. Loudcher, "A joint model for topic-sentiment evolution over time," in 2014 IEEEInternational

Conference on Data Mining, Dec 2014, pp. 773–778.

6. D Tang, B. Qin, and T. Liu, "Learning semantic representations of users and products for document level sentiment classification," in Proceedings of the 53rd Annual Meeting of the Association forComputational Linguistics and the 7th International Joint Conferenceon Natural Language Processing (Volume 1: Long Papers). Beijing, China: Association for Computational Linguistics, July 2015, pp. 1014–1023.
7. X Hu, L. Tang, J. Tang, and H. Liu, "Exploiting social relations for sentiment analysis in microblogging," in Proceedings of the SixthACM International Conference on Web Search and Data Mining, ser. WSDM '13. New York, NY, USA: ACM, 2013, pp. 537–546.
8. S Liu, X. Cheng, F. Li, and F. Li, "Tasc:topic-adaptive sentiment classification on dynamic tweets," IEEE Transactions on Knowledgeand Data Engineering, vol. 27, no. 6, pp. 1696–1709, June 2015.
9. M M. Rahman and H. Wang, "Hidden topic sentiment model," in Proceedings of the 25th International Conference on World Wide Web, ser. WWW '16. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2016, pp. 155–165.
10. G. Paltoglou and M. Thelwall, "A study of information retrieval weighting schemes for sentiment analysis," in Proceedings of the48th Annual Meeting of the Association for Computational Linguistics, ser. ACL '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 1386–1395