Sentimental Analysis on Live Tweets

K.Karpaga priyaa	V.Ashwini	K.Gayathri	S.Shweta
Department of CSE	Department of CSE	Department of CSE	Department of CSE
Sri Sairam	Sri Sairam	Sri Sairam	Sri Sairam
Engineering College	Engineering College	Engineering College	Engineering College
Chennai	Chennai	Chennai	Chennai

ABSTRACT

Social networking has become a very powerful communicative tool among internet users. Apart from communication millions of users share their opinion on everyday issues. Twitter, for instance, offers multiple mechanisms for digital communication, including drop-down commenting and scheduled support group discussion. The event based discussion on twitter can be analysed to find the opinion of people towards a particular event. Sentimental analysis can be performed on tweets to derive a conclusion about people's opinion. The main problem in the existing approaches are, in the event based discussion, only static data is used for analysis of public emotions which results in incomplete analysis of an event . The dynamic live tweets are also equally important on live events like sports, election and so on. Thus the proposed systems uses dynamic unused raw data existing in twitter and produces the polarity of live comments using the classification algorithm. And also this paper takes into account, the sarcastic tweets on twitter in particular, with measurable data, available to public domain.

1.INTRODUCTION

Social media is steadily impacting the way we live and communicate with one another. It serves as a valuable platform for people to share, discuss, and engage in social issues that are important to the community. Twitter, in particular, has become one powerful tool for marketing and communication[2],[9],[3],[8]. Participants in a twitter conversation can be as an observer (using 'Like' option)[10], criticizer (using 'Retweet' option) or commenter (using 'Reply' option). These tweet opinions mainly talk about organizations and individuals who have a strong social, political or economic interest in maintaining and enhancing their popularity and trustworthiness. Sentiment analysis enables these organizations the ability to monitor different social media sites in real time[18],[14],[15].

It helps in determining the emotional texture behind words to gain an understanding of the attitudes and emotions expressed within an online mention. The attitude analyzed by sentimental analysis may be a judgment or evaluation, affective state, or the intended emotional communication. This technology is particularly used to collect and analyze opinions about product, service or even a brand on Twitter. It mainly works on gaining an overview of the wider public opinion behind certain topics and emotional reaction to a document, interaction or event. The main intention is to check whether a piece of writing is positive or negative. The assortment done by sentimental analysis(SA) is through continuously exploiting the static data on a particular domain and furnishing the final outlook of the tweets[2]. The analysis is done through several steps which include, textual preprocessing (tokenization, stemming, lemmatization, noise rectification)[7], classification and determining polarity using a supervised algorithm of naive bayes [1] on training data ranges. The major problem faced by the existing system is, only the static tweet analysis of an event is taken into consideration even if the issue is a present-day domain under discussion.

In collateral to this, the system fails if the raw unstructured tweet conveys both positive and negative sense. Another serious challenge faced in the current system is, the misspelling of tweets and Twitter spamming[4],[17].

Hence, we approach this drawback with our proposed simple idea of taking into account the dynamic tweets along with the static tweets and providing a satisfactory conclusion to the users. The method extends through including another training data set for identifying both positivity and negativity in order to analyse the neutral tweets. The proposed system supplies a spell checking API for correcting the misspelled tweets. Spamming on twitter is rectified through checking authorization and capturing the repetition of words by a training data set. The multinomial naive Bayes algorithm includes

proper training data sets prior to detect the polarity of the tweets.

In addition the supervised model[14],[18] also analysis sarcastic tweets and categorizes them into positive, negative and neutral.

2.LITERATURE SURVEY

Sentiment Analysis

Sentiment analysis helps to extract the emotion from a statement. The main contribution in the field of sentiment analysis focus classifying the given text into different classes i.e positive or negative [1],[10],[14],[18]. The paper[5] classifies the text into more classes and provides seven different sentiment classes in twitter. The work on[7] focus on various tweet preprocessing techniques and it proves that Naïve Bayes and Random Forest classifiers are more sensitive than Logistic Regression and Support Vector Machine.[1] Used Naïve Bayes classifier for performing sentimental analysis on the contents in newspaper,e-book, etc and compresses it to bring forth pictorial representation.[2] Uses n-gram features and word sentiment polarity integrated with a deep convolutional neural network to analyze the twitter data. The paper [3],[6] discusses the challenges faced in the analysis of dynamic data and the impact of live tweets on live movements.[8] Performs survey and experimental analysis on event detection techniques for twitter data streams.[4] Focus on event-based discussion by obtaining the tweets related to a particular event.

Tweets

Most existing works on twitter sentiment analysis uses supervised methods. Supervised methods are based on classifiers such as Naïve Bayes , Support Vector Machine, Random Forest and so on[5],[6],[7] .The tweets are classified based on various combination of features such as part of speech, emotion, tweet context etc[7],[9]. Other than supervised methods unsupervised and deep learning methods are also used .Deep learning is a powerful machine learning technique which learns the data from multiple layers and produces the result[2],[11]. The deep learning neural networks are initially used in voice and speech recognition ,apart from its implementation in many other machine learning domains, it is also used in sentiment analysis. Though deep learning methods produce satisfactory results, it is more complicated as the architecture have to be fine-tuned to achieve best performance, it

needs lots of data as well as long training times. thus using supervised methods gives optimal solution[7].

If we take a step into event based discussions in twitter ,lots of people will tweet their opinion about the event .If we consider events such as live sports match, election[6] lots of people will tweet with the particular hashtag and the opinion of people will change by time to time since scenarios in a live event changes, for example in a public election the people support a candidate changes from day to day. Thus it is important to capture the live happening of an event .The existing approaches use sentiment analysis to classify the tweets but most of the approaches uses only static data[1],[2],[5],[4]i.e the static tweets tweeted by the users .Using only the tweets which are tweeted earlier (which may be tweeted before an hour or day or day before yesterday) does not helps in the proper analysis of the event .This results in the incomplete collection of data about an event , which leads to wrong judgement about the event .Thus collecting the live tweets of an events will only help in deciding the public emotion.

Sarcastic Comments

There may be many tweets with indirect meaning such as metaphor ,sarcastic tweets etc[6],[3]. The analysis of sarcastic tweets is also important in providing public opinion .Though several neural networking techniques[2],[11] are used, the analysis of sarcastic tweets is still challenging in the existing approaches..

The main drawback in the existing approaches are they use only static data for analysis and sarcastic comments are not given much importance,

3.PROPOSED SYSTEM

Since the significance of live data in an event based discussion is high, the proposed system uses live tweets for analysis. The system collects the tweets from the live stream which are related to a particular topic using keyword matching technique. Sentiment analysis is performed on the retrieved live tweets using classification algorithm .Multinomial Naive Bayes classification algorithm is used to classify the tweets into positive and negative[1] .We have tested several algorithms by performing sentiment analysis on static data, the results shows that Naïve Bayes algorithm shows greater percentage of accuracy than other supervised learning algorithms .In the Naïve Bayes algorithms, Multinomial Naïve Bayes algorithm is used since it is more appropriate for

text-based analysis. The proposed system also concentrates on removing the spam tweets[4],[17] .Spamming is an challenging and important problem since the spam tweets may contain infectious URL[4] ,inappropriate collection of user's data by asking the user to give some information .Spamming also disturbs the opinion mining process since a single user may tweet for several times(several tweets of single user should be taken into account as one).This spamming problem is solved by using specific training data set which contains the words mostly used by spammers and it filters the several tweets tweeted by single user by tracking the user id. A separate classification for neutral tweets[5] is also made along with the positive and negative.

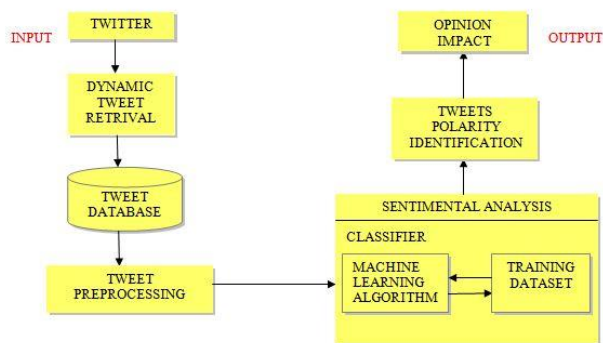Fig. 1 represents the actual implementation of the system.

Fig. 1 Architectural Diagram

Dynamic tweet retrieval:

 The live tweets are collected from the twitter using twitter's API. The sign up for twitter stream API is made by using a twitter account[9]. The stream API retrieves the live tweets, the search results can be restricted by using the appropriate search query. The inappropriate results which includes lots of page name and profile name can be filtered with the help of the API itself.

Tweet preprocessing:

There are several pre-processing techniques in the existing approaches. The essential techniques like stopword removal, hashtag removal, URL removal, extra symbol removal are performed[7]. Additional to this grammar parsing and spell checking is done to map with the proper word in the training data set.

Sentiment analysis:
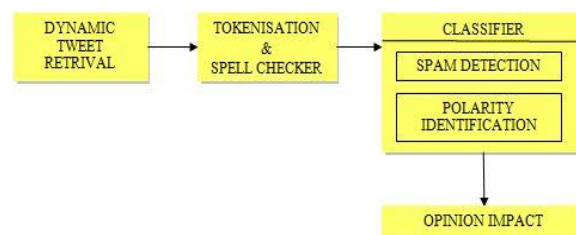
In Fig. 2 the submodules are explained below.

Fig. 2 Sub modules of Sentimental analysis

A. Spam detection:

 The spam tweets are filtered by training the algorithm using specific training data set [4],[17]. The training data set contains the words which are used by spammers. The tweets are checked for any spam words, if the spam words are present in the tweet then it checks for the number of spam words. If the presence of spam words are higher the normal words then the algorithm classifies it as spam and removes the particular tweet. The same word mapping technique is used to find single user with several tweets.

B. Tweet polarity:

 Multinomial Naïve Bayes algorithm is used to classify the tweets[1]. The training data set used for classification contains several words, if a new word arises in a tweet,a search for the meaning of the word is made through data dictionary and the word is updated in the training data set. We use three set of training data sets for classification first is positive data set, second is negative data set and third data set contains both positive and negative words in pairs, which are the combinations present in most sarcastic and neutral comments i.e the pattern observed in the neutral tweets and sarcastic tweets are combined as a set and are stored in the training data set. Using these three data set the tweets are classified as positive, negative and neutral.

C. Classifier

I. Algorithm

Step 1: Input ← Live_tweets {Get the input as live tweets}

Step 2: Preprocess the tweets
    i)   Tokenize
    ii)   Spell check

$\alpha, \beta, \gamma, \ldots \leftarrow$ data1, data2, data3 ,....

$S_\alpha, S_\beta, S_\gamma, \ldots \leftarrow \alpha, \beta, \gamma, \ldots$

where,

$\alpha, \beta, \gamma, \ldots$ are tokenized terms

$S_\alpha, S_\beta, S_\gamma, \ldots$ are spell-checked terms

Step 3: Spam detection

Multinomial Naïve Bayes Classifier is used for spam detection.
Call method multi_naivebayes_classifier(S_α, S_β, S_γ, ...) in a recursive way.
Multinomial Naïve Bayes Classifier is formulated as,

$$\hat{\theta}_i = \frac{x_i + \alpha}{N + \alpha d} \qquad (i = 1, \ldots, d),$$

where,

$x = (x1, x2, \ldots, xd)$ is the observation from a multinomial distribution with N trials
$\theta = (\theta1, \theta2, \ldots, \theta d)$ is the parameter vector

α is smoothing parameter, $\alpha > 0$

Step 4: Classify tweets into positive and negative terms.
Call method multi_naivebayes_classifier(S_α, S_β, S_γ, ...) in a recursive way.

Step 5: Cluster the polarity of tweets to produce the polarity in terms of percentage.
    K-means clustering algorithm is used for clustering.
    K-means clustering algorithm is formulated as,

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 ,$$

where,

$\| xi(j) - cj \|2$ is a chosen distance measure between a data point $xi(j)$ and the cluster centre cj .

Step 6: Output $\leftarrow$ Polarity of Live_tweets (in %){
Polarity of live tweets is produced as output in terms of percentage (%) }

Ii. Experimental Evaluation

On performing spam detection on movie reviews which is static data, the accuracy of Multinomial Naïve Bayes Classifier and Bernouille Naïve Bayes Classifier is as follows:

CLASSIFIER	ACCURACY
Multinomial Naïve Bayes Classifier	0.761
Bernouille Naïve Bayes Classifier	0.755

With the above proof, Multinomial Naïve Bayes Classifier is chosen to perform spam detection on live tweets i.e., dynamic data.

Iii. Graphical Representation

The graphical representation of the accuracy of both the classifiers on performing sentimental analysis on static data is as follows:
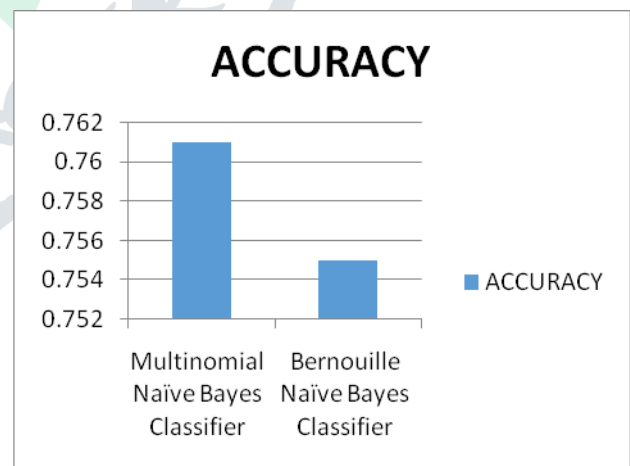
Tweet polarity identification:

After the words in the tweets are classified as positive and negative the number of positive words and negative words in a tweet in calculated to tag a particular tweet as positive or negative. Then an aggregate value is calculated to find the total public opinion about the particular event.

Opinion impact:

For a particular event the percentage of people tweeted positive ,negative ,neutral is produced. Thus we can able to know the total opinion in the social media.

4.ADVANTAGES

Today's world is the customer's world. The feedback and review of customers are important to develop the business. This analysis of customer reviews can be automated using sentiment analysis techniques. Apart from the customer reviews, social networking site can be used to know the people opinion about several products, brands, current happenings[2],[18],[14]. The proposed system goes to the fullest form of analysis of social media trends by taking the streaming tweets from Twitter

The system is useful for both enterprise and also common users. The enterprise owners will be able to know what people are talking about their product on Twitter. The common users can know the opinion of people about trending news, a live event or a trending hash tag.

As live tweet differs widely from regular, everyday tweeting, it requires participants to be focused on the event hash tag. So anything we tweet during this time has the potential to get attention and make an important impact. Therefore, live tweet analysis is a great benefit of building a confidential opinion of targeted users and it also increases the accuracy of finding the opinion of the audience. It helps in opinion mining of any targeted site, generates interest for the content or to make an important announcement during an event. It also builds interest from people to tweet about any product and services to build a relationship with influencers. Thus, dynamic sentimental analysis generates a wholesome opinion by taking into account both static and dynamic data in an event-based discussion

Certain tweet contains neither positive nor negative words. These texts can be considered as neutral. Neutrality of comments other than positive and negative is taken into consideration in our work. The introduction of the neutral category improves the overall accuracy of the system.

The spell checker on sentimental analysis increases the accuracy of the system. It encourages ease in typing on the computer and decreases number typos in the document.

Spam detection on sentimental analysis drastically rules out the scammers from posting unrelated tweets, posting duplicate updates and providing links to the website.

Identification of tonal and gesture clues in the textual data is very difficult. Wherefore opinion analysis evaluates these sarcastic tweets and finds their polarity using different training data sets.

5.CONCLUSION

In this paper, we present dynamic tweet analysis: the mechanism which is helpful to analyze the information of the real-time tweets where opinions are highly unstructured, divergent and are either positive or negative or neutral in some cases. The opinion mining in our discussion takes into account both static and dynamic tweet analysis and draws on a potent conclusion at the end. On contrary to the existing system, the live analysis is performed by using the event hash tag to prompts the follower's interest to interact on a particular event discussion.

The accuracy of the system has been uplifted by including a spell checker and spam detection. Moreover, the analysis of proposed system extends its arm in determining the polarity of sarcastic tweets. Throughout this work, we demonstrated that the Sentimental analysis of dynamic tweets can achieve a high level of accuracy and capture exactly the preciseness of the opinion. Future works focus on predicting the opinion of emotion symbols like emoji ideograms, and for analysis of memes and metaphors on Twitter

7.REFERENCES

[1] A Akshai R ,B Meenakshi S,C Rohit Krishnan ,D. S Shweta and E Venkatesh B P, "Editorial Context Solution Using Real Time Interactive Visualization-Data Driven Story" Proceedings of International Journal of Information and Electronics Engineering vol. 8,no 3, pp 36-40,2018.

[2] Zhao Jianqiang , Gui Xiaolin, and Zhang Xuejun1 "Deep Convolution Neural Network For Twitter Sentiment Analysis" Proceedings of NSFC projects, vol 6, pp 1-7,2018.

[3] Brandeis Marshall, Takeria Blunt, Tayloir Thompson "The Impact Of Live Tweeting On Social Movements " Proceedings of IEEE International Conference on Information Reuse and Integration for Data Science, pp 1-10,2018.

[4] Koichi Sato , (Member, IEEE), Junbo Wang , (Member, IEEE), And Zixue Cheng, (Member,

IEEE) "Credibility evaluation of twitter based event detection by a mixing analysis of hetrogeneous data" Proceedings of JST, Strategic International Collaborative Research Program, SICORP ,vol 7, pp 1-11,2018.

[5] Mondher Bouazizi And Tomoaki Ohtsuki "A Pattern-Based Approach for Multiclass Sentiment Analysis in Twitter" Proceeding of IEEE International conference ,vol 7,pp 1-21,2017

[6] Monireh Ebrahimi, Amir Hossein Yazdavar, And Amit Sheth, Kno.E.Sis Center "Challenges of Sentiment Analysis for Dynamic Events" Proceedings of IEEE computer society, pp 1-4, 2017.

[7] Zhao Jianqiang , Gui Xiaolin, "Comparison Research on Text Pre Processing Methods on Twitter Sentiment Analysis." Proceedings of IEEE computer society ,vol 5,pp 1-9,2017.

[8] Andreas Weiler , Michael Grossniklaus and Marc H. Scholl "Survey and Experimental Analysis of Event Detection Techniques for Twitter" Proceedings of British Computer Society ,vol 60,Isuue 3, pp 329-346,2016.

[9] N. F. F. da Silva, E. R. Hruschka, and E. R. Hruschka, ''Tweet sentiment analysis with classifier ensembles,''Proceedings of Decision Support Syst., vol. 66, pp. 170–179, Oct. 2014

[10] B. Pang, L. Lee, and S. Vaithyanathan, ''Thumbs up?: Sentiment classification using machine learning techniques,'' Proeedings of. Conf. Empirical Methods Natural Lang. Process.,pp. 79–86,2002.

[11] C. dos Santos and M. Gatti, ''Deep convolutional neural networks for sentiment analysis of short texts,'' Proceedings of 25th Int. Conf. Computer. Linguistics, Dublin, Ireland, pp. 69–78,2014.

[12] R. Dayani, N. Chhabra, T. Kadian, And R. Kaushal, ''Rumor Detection In

Twitter: An Analysis In Retrospect,'' Proceedings of Ieee Int. Conf. Adv. Netw.

Telecommun. Syst, pp. 1–3,2014.

[13] J. Schmidhuber, "Deep Learning In Neural Networks: An Overview,"

Corr, Vol. Abs/1404.7828, 2014. [Online]. Available: Http://Arxiv.Org/

Abs/1404.7828 ,2014.

[14] H. Saif Et Al " Senticircles For Contextual And Conceptual Semantic Sentiment S. M. Kim and E. Hovy, ''Determining the sentiment of opinions,'' Procdings of

20th Int. Conf. Comput. Linguistics Assoc. Comput. Linguistics, p. 1367,2015.

[15] H. Saif, Y. He, M. Fernandez, and H. Alani, ''Adapting sentiment lexicons using contextual semantics for sentiment analysis of twitter,'' Proceedings of Eur. Semantic Web Conf, pp. 54–63,2014.

[16] M. Wang and H. Shi, ''Research on sentiment analysis technology and

polarity computation of sentiment words,'' Proceedings of IEEE Int. Conf. Prog.

Informat. Comput. (PIC), pp. 331–334,2010

[17] X. Hu, L. Tang, J. Tang, and H. Liu, ''Exploiting social relations for

sentiment analysis in microblogging,'' in Proc. 6th ACM Int. Conf. Web

Search Data Mining, pp. 537–546,2013

[18] S. M. Kim and E. Hovy, ''Determining the sentiment of opinions,'' in Proc.

20th Int. Conf. Comput. Linguistics Assoc. Comput. Linguistics,

pp 150-160,2004..