Sarcasm Detection and User Behaviour Analysis

Pooja Deshmukh
Student of ME (CSE)
Department of Computer Science and Engineering
Deogiri Institute of Engineering and Management
Studies, Aurangabad

Sarika Solanke
Assistant Professor
Department of Computer Science and Engineering
Deogiri Institute of Engineering and Management
Studies,Aurangabad

Abstract**—**Sarcasm is a sort of sentiment where public expresses their negative emotions using positive word within the text. It is very tough for humans to acknowledge. In this way we show the interest in sarcasm detection of social media text, particularly in tweets. In this paper we propose new method pattern based approach for sarcasm detection, and also used behavioral modelling approach for effective sarcasm detection by analyzing the content of tweets however by conjoint exploiting the activity traits of users derived from their past activities. In this way we propose the different method for sarcasm detection such as, Sentiment-related Features, Punctuation-Related Features, Syntactic and Semantic Features, Pattern-Related Features approach for detection of sarcasm in the tweet. We also develop the behavioural modeling approach to check the user emotion and sentiment analysis. By using the various classifiers such as TREE, Support Vector Machine (SVM), BOOST and Maximum Entropy, we check the accuracy and performance. Our proposed approach reaches an accuracy of 94 %.

Keywords*-Sarcasm, Sentiment, SVM, BOOST.*

*****

I. INTRODUCTION

Social net-working websites have become a popular platform for users to express their feelings and opinions on various topics, such as events, or products. Social media channels have become a popular platform to discuss ideas and to interact with people worldwide area. Twitter is also important social media network for people to express their feelings, opinions, and thoughts. Users post more than 340 million tweets and 1.6 billion search queries every day [1] [2].

Twitter is a social media platform where users post their views of everyday life. Many organizations and companies have been interested in these data for the purpose of studying the opinion of people regards the political events, popular products or Movies. When a particular product is launched, people start tweeting, writing reviews, posting comments, etc. on social media such as twitter. People turn to social media network to read the comments, and reviews from other users about a product before they decide whether to purchase or not. If the user review is good for the particular products then the users are buy the product otherwise not. Organizations are also depends on these sites to know the response of users for their products and use the user feedback to improve their products [3]. Sentiment analysis is the opinion of the user for the particular things. Sentiment analysis is the extraction of feeling from any communication (verbal/non verbal).Two ways to express sentiment analysis.

1) Explicit sentiments: Direct expression of the opinion about the subject shows the presence of explicit sentiment.
2) Implicit sentiments: Whenever any sentence implies an opinion then such sentence shows the Presence of implicit sentiment (Indirect expression).

Sentiment analysis and opinion mining depends on emotional words in a text to check its polarity (i.e., whether it deals positively or negatively with its theme) [4].Sarcasm is a type of sentiment where people express their negative feelings using positive word in the text. The example of this is “I love the pain of breakup”. The love is the positive words but it expresses the negative feeling, such as breakup in this example. It is usually used to transfer implicit information within the message a person transmits. It is hard even for humans to recognize. Used Pattern based approach for detecting sarcasm on twitter. The definition of sarcasm is the activity of saying or writing the opposite of what you mean, or of speaking in a way intended to make someone else feel stupid or show them that you are angry. Also check the user behaviour, it used for sarcasm detection.

II. LITERATURE REVIEW

In [3], authors show the interest in sarcasm delectation in the tweeter. For capturing real time tweets they use the Hadoop base framework, and processes that tweets they used the different six algorithms such as parsing based lexicon generation algorithm (PBLGA), tweets contradicting with universal facts (TCUF), interjection word start (IWS), positive sentiment with antonym pair (PSWAP), Tweets contradicting with time-dependent facts (TCTDF), Likes dislikes contradiction (LDC), these algorithm are used identifies sarcastic sentiment effectively. This method is more suitable for real time streaming tweets.

In [4], authors use the computational system it is use for harnesses context incongruity as a basis for sarcasm detection. Sarcasm classifier uses four types of features: lexical, pragmatic, explicit incongruity, and implicit incongruity features. They evaluate system on two text forms: tweets and discussion forum posts. For improvement of performance of tweet uses the rule base algorithm, and to improve the performance for discussion forum posts, uses the novel approach to use elicitor posts for sarcasm detection. This system also introduces error analysis, the system future work (a) role of numbers for sarcasm, and (b) situations with subjective sentiment.

In [5], authors used the machine learning approach to sarcasm detection on Twitter in two languages English and Czech. First work is sarcasm detection on Czech language. They used the two classifier Maximum Entropy (MaxEnt) and Support Vector Machine (SVM) with different combinations of features on both the Czech and English datasets. Also use the different preprocessing technique such as Tokenizing, POS-tagging, No stemming and Removing stop words, its use for finding the issue of Czech language.

In [6], authors have investigated characteristics of sarcasm on Twitter. They are concerned not just with identifying whether tweets are sarcastic or not, but also consider the polarity of the tweets. They also have compiled a number of rules which improve the accuracy of sentiment analysis when sarcasm is known to be present. Resercher have developed a hash tag tokenizes for GATE method so that sentiment and sarcasm found within hash tag can be detected more easily. Hash tag tokenization method is very useful for detection of sarcasm and checks the polarity of the tweet i.e. positive or negative.

In [7], authors are used two methods such as lexical and pragmatic factors that are use for differentiate between sarcasm from positive and negative sentiments expressed in Twitter messages. They also created corpus of sarcastic Twitter messages in which determination of the sarcasm of each message has been made by its author. Corpus is used to compare sarcastic utterances in Twitter to utterances that show positive or negative attitudes without sarcasm.

In [8], authors have developed a sarcasm recognizer to determine sarcasm on Twitter consists of a positive sentiment contrasted with a negative situation of sarcasm in tweets. They use novel bootstrapping algorithm that automatically learns lists of positive sentiment phrases and negative situation phrases from sarcastic tweets. They show that determine contrasting contexts using the phrases learned through bootstrapping.

Rule-based approaches attempt to identify sarcasm through specific evidences. These evidences are captured in terms of rules that rely on indicators of sarcasm. Focus on identifying whether a given simile (of the form '* as a *') is intended to be sarcastic. They use Google search in order to determine how likely a simile is. They present a 9-step approach where at each step rule; a simile is validated using the number of search results. Strength of this approach is that they present an error analysis corresponding to multiple rules [9].

The hash tag sentiment is a key indicator of sarcasm. Hash tags are often used by tweet authors to highlight sarcasm, and hence, if the sentiment expressed by a hash tag does not agree with rest of the tweet, the tweet is predicted as sarcastic. They use a hash tag tokenizer to split hashtags made of concatenated words [6].

III. SYSTEM ARCHITECTURE

In this work, we propose two approaches i.e. sarcasm detection based and behavioral modeling approach.A pattern-based approach to detect sarcasm on Twitter. Propose four sets of features that cover the different types of sarcasm we defined. We use those to classify tweets as sarcastic and non-sarcastic [11]. Also used behaviour modelling approach to develop a systematic approach for effective sarcasm detection by not only analyzing the content of the tweets but by also exploiting the behavioral traits of users derived from their past activities [15].

1) Sarcasm Detection System

The architecture of proposed system is shown in Fig 1. We have developed the sarcasm detection system with pattern based approach.

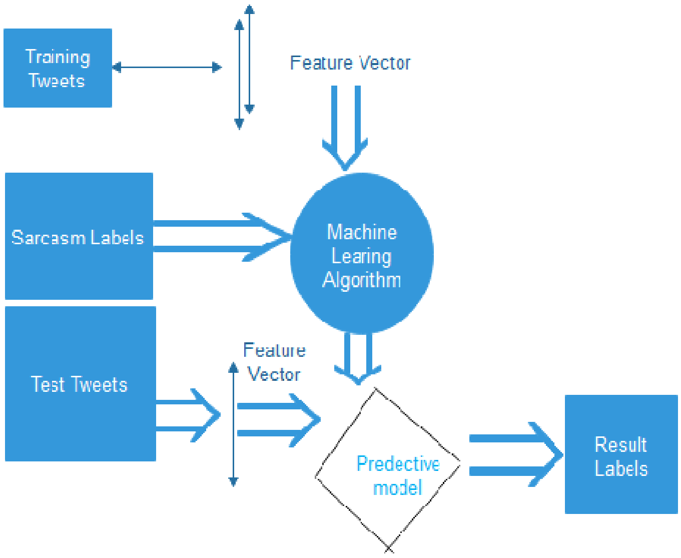


Fig 1 System Architecture of Sarcasm detection

The above architecture shows the working of the sarcasm detection system.

1) Training tweets:

ISSN: 2321-8169
77 - 82

The Training tweets contain the 5000 tweets are collected by using tweeter API. The collected tweets are a list format converted into the csv (comma separated word) format.

2) Feature Vector or Features extraction:

Four types of Feature are extracted. This method are used for annotating the data, it contain three categories.

a) Sarcasm as wit: when used as a wit, sarcasm is used with the purpose of being funny.

b) Sarcasm as whimper: when used as whimper, sarcasm is employed to show how annoyed or angry the person is.

c) Sarcasm as evasion: it refers to the situation when the person wants to avoid giving a clear answer, thus, makes use of sarcasm.

i) Sentiment-related Features

It extracts sentimental components of the tweet and counts them. Positive emotional content (e.g. love, happy, etc.) and negative emotional content (e.g. hate, sad, etc.).Calculate the ratio of emotional words.

p (t) = (& · PW + pw) − (& · NW + nw)/ (& · PW + pw) + (& · NW + nw) 1

t=tweet, pw=positive words, nw =negative words, PW=highly emotional positive words, NW= highly emotional negative words, & =weight bigger than 1.

ii) Punctuation-Related Features

It displays behavioral aspects such as low tones, Facial gestures or exaggeration. These aspects are translated into a certain use of punctuation or repetition of vowels when the message is written.

- Number of exclamation marks
- Number of question marks
- Number of dots
- Number of all-capital words
- Number of quotes

iii) Syntactic and Semantic Features

It refers to the situation when the person wants to avoid giving a clear answer, thus, makes use of sarcasm.

- Use of uncommon words
- Number of uncommon words
- Existence of common sarcastic expressions
- Number of interjections
- Number of laughing expressions

iv) Pattern-Related Features

Pattern is defined as an order sequence of words. Divide words into two classes: a first one called as CI containing words of which the content is important and a second one called to as GFI containing the words of which the grammatical function is more important.

Step to develop pattern based approach.

1) Take the tweet
2) POS tag
3) Pattern Extraction
4) Tokenization
5) Count frequency of pattern
If frequency = 2 then
Add the pattern otherwise discards the pattern
6) Calculate resemblance degree

- *res(p, t)*

$$\begin{cases} 1 & \text{if the tweet vector contains the pattern as it is, in the same order;} \\ ^{\delta}\,.n/N; & \textit{if n words out of the N words of the pattern} \text{ appear in the tweet in the correct Order;} \end{cases}$$

0, if no word

2

7) Calculate feature set

$$\mathrm{Fij} = \beta j \sum_{k=0}^{k} \mathrm{res(Pk, t)}$$

3

Where Bj is a weight given to patterns of length *Lj* is their level of sarcasm. *Fij* is calculate the degree of resemblance of a tweet *t* to patterns of level of sarcasm *i* and length *j*. K in our work is set to 5, and represents the K closest patterns among the *Nij*.

3) Sarcasm label:

The sarcasm labels are also provided i.e. 0 to 5 mean 0, 1, 2, 3, 4, 5.the training data labels as sarcasm labels and it passes to the machine leaning algorithm.

4) Machine learning algorithm

The Supervised learning algorithms are used.

Following machine learning algorithm are used.

a) MaxEntropy
b) SVM
c) Tree
d) Boost

5) Test Tweets:

The 1000 testing tweets are available to test the machine learning result. If the machine learning and testing tweets give the same result then our approach is giving good accuracy.

6) Predictive modelling:

The machine learning and testing tweets result are comparing in the predictive modelling. Finally we get the accurate result label. In this way the sarcasm detection architecture is work.

2) Behavioural modelling approach

The second approach is user behavioural modelling .To develops a systematic approach for effective sarcasm detection by not only analyzing the content of the tweets but by also exploiting the behavioral traits of users derived from their past activities this system is used. Following are the features

a) Hashtag used by or for user
b) Word used by or for user
c) Positive word used by or for user
d) Negative word used by or for user

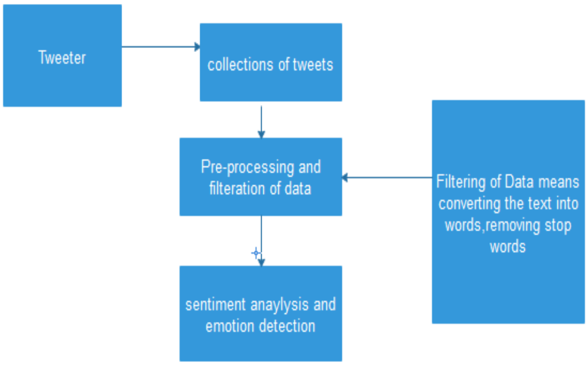

Tweeter
collections of tweets
Pre-processing and filteration of data
Filtering of Data means converting the text into words,removing stop words
sentiment anaylysis and emotion detection

Fig 2 System Architecture of Behavioural modeling

1) Tweeter:

Tweeter is the social media network, which is use for communication. Also used for share the opinions for the user throw the tweets. A tweet is collected by using tweeter API. The 1000 tweets are collected.

2) Pre-processing and filtration of data

Many current methods for text sentiment analysis contain various preprocessing steps of text. One of the most important goals of preprocessing is to enhance the quality of the data by removing noise. Another point is the reduction of the feature space size.

3) Sentiment analysis and emotion detection:

After the preprocessing of the data the next step is the sentiment analysis and user emotion detection. User behavioral is very important to check the user emotion. Emotion detection contains the emotion of the user like happy, angry, joy etc. Check the user emotion using their past tweets. This is the workings of the behavioural modelling approach.

IV. PERFORMANCE EVALUATION

We have evaluated the performance of our proposed system. In this section, we present experimental results on Sarcasm detection & behavioral modeling approach and increase in result accuracy, efficiency.

The Key Performance Indicators (KPIs) used to evaluate the approach are:

1) Accuracy: it represents the overall correctness of classification. In other words, it measures the fraction of all correctly classified instances over the total number of instances.

2) Precision: it represents the fraction of retrieved sarcastic tweets that are relevant. In other words, it measures the number of tweets that have successfully been classified as sarcastic over the total number of tweets classified as sarcastic.

3) Recall: it represents the fraction of relevant sarcastic tweets that are retrieved. In other words, it measures the number of tweets that have successfully been classified as sarcastic over the total number of sarcastic tweets.

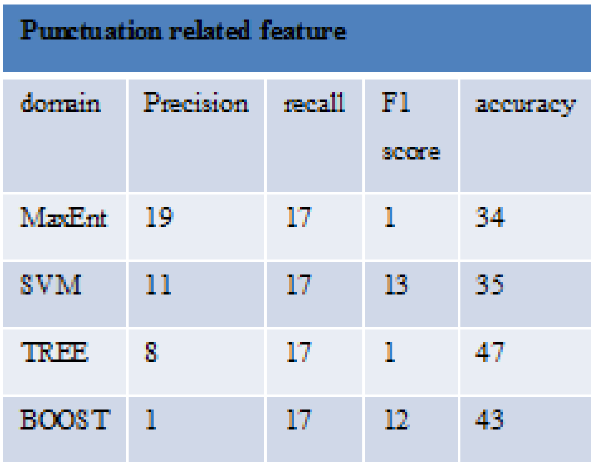
4) F1 score:

F1 =2 * (precision * recall/precision + recall)

1) Results

The following section presents results of all the experiments discussed in Table, and graph. All the experiments results are shown feature wise, i.e. the result of four experiments is shown for Punctuation related firstly, then sentiment, syntactic and lastly Pattern based. Then behavioral modeling result is shown.

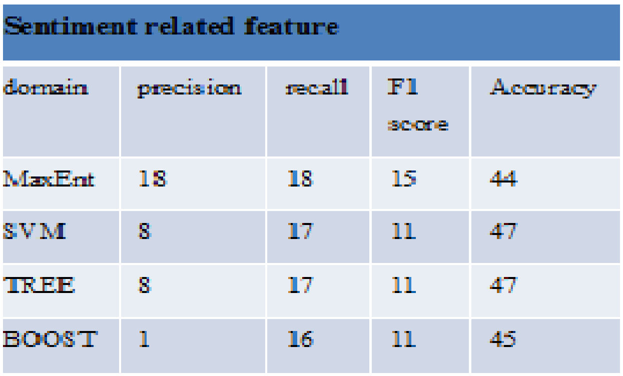
Below table shows the result of four feature methods using the different algorithm. Test Result Set for Feature Extraction Methods



Punctuation related feature

domain	Precision	recall	F1 score	accuracy
MaxEnt	19	17	1	34
SVM	11	17	13	35
TREE	8	17	1	47
BOOST	1	17	12	43

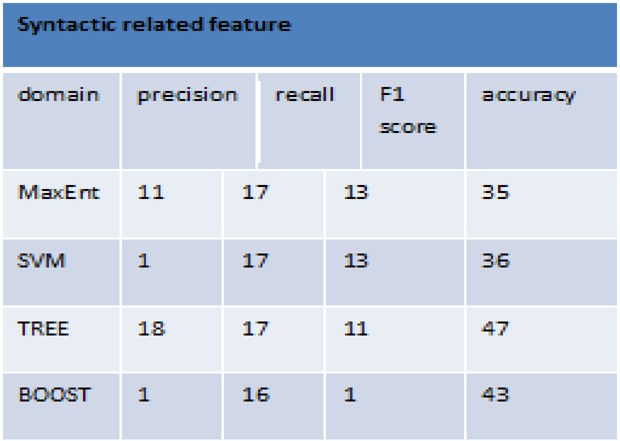
Table (a)



Sentiment related feature

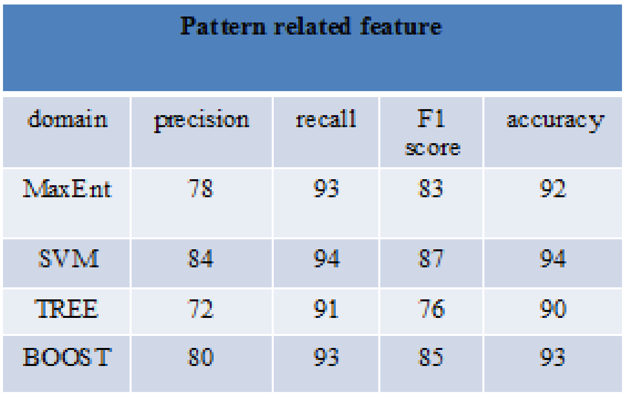
domain	precision	recall	F1 score	Accuracy
MaxEnt	18	18	15	44
SVM	8	17	11	47
TREE	8	17	11	47
BOOST	1	16	11	45

Table (b)



Syntactic related feature				
domain	precision	recall	F1 score	accuracy
MaxEnt	11	17	13	35
SVM	1	17	13	36
TREE	18	17	11	47
BOOST	1	16	1	43

Table (c)



Pattern related feature				
domain	precision	recall	F1 score	accuracy
MaxEnt	78	93	83	92
SVM	84	94	87	94
TREE	72	91	76	90
BOOST	80	93	85	93

Table (d)

The Above table shows the result of the four features using the different algorithm. Features are sentiment, punctuation, syntactic and pattern related feature. The pattern Based feature give the more result as compare to other three features, the pattern based gives the highest accuracy i.e. 94%.Pattern based is used for sarcasm detection, the result are calculated by using the different classifiers, the classifiers are SVM(support vector machine),TREE, BOOST, MaxEnt. Following are the Graphical Representation of Experimental Results on four feature sets.

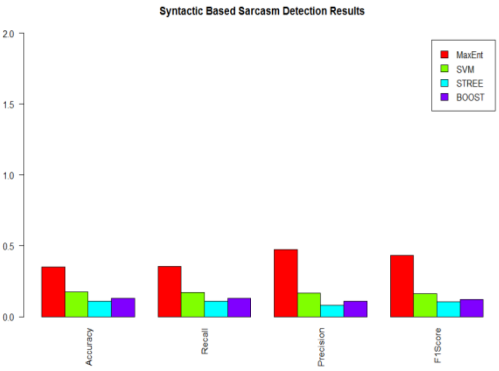


Fig (a)

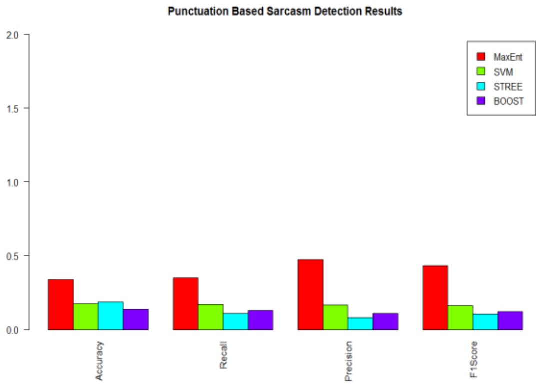


Fig (b)

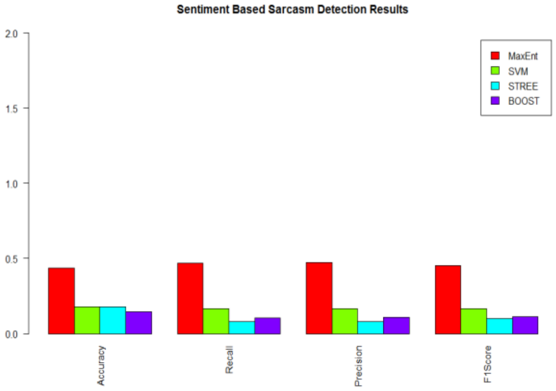


Fig (c)

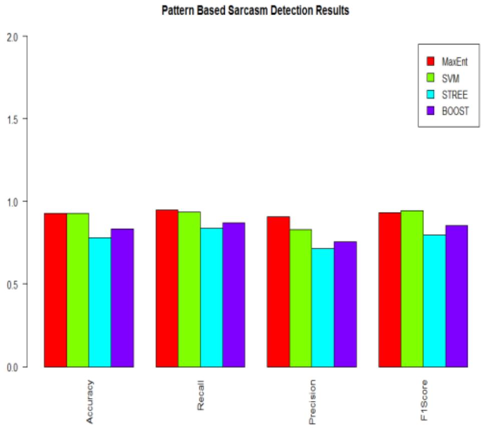


Fig (d)

Behavioral analysis

Here we have shown some old twits real time user behavioral analysis

The user is considering as most popular person, for example Mr. nfl, the following graph showing such analysis based on his twits.

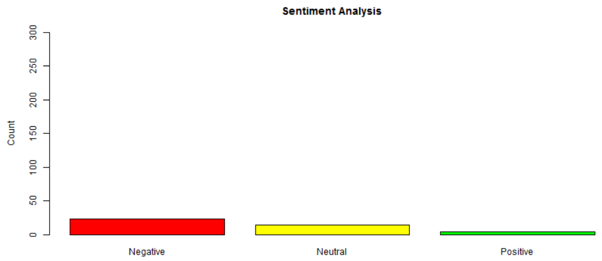


Fig (e)

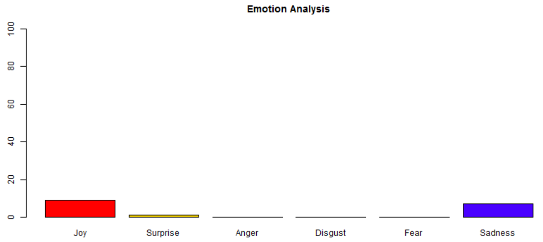


Fig (f)

V. CONCLUSION AND FUTURE WORK

In this paper, the proposed methods are used to detect sarcasm or as well as check the behavioral approach of the user, the method make used different component of the tweet, and also by using of Part-of-Speech tags to extract patterns characterizing the level of sarcasm of tweets.We collect the all sarcastic tweets by using #sarcasm.In this way we implemented the different method for sarcasm detection such as, Sentiment-related Features, Punctuation-Related Features, Syntactic and Semantic Features, Pattern-Related Features approach for detection of sarcasm in the tweet as compare to all methods the pattern related feature gives more result. Behavioural modelling approach for detection of sarcasm in the tweet. Behavioral modeling used to check the emotion, and sentiment analysis for the user.The naïve bayes classifier is used to check the emotion and sentiment analysis of the use. By using different algorithm or classifier such as BOOST, Support Vector Machine (SVM), TREE and Maximum Entropy, check the accuracy and performance. Proposed method gives more result as compare to previous. Our proposed approach reaches an accuracy of 94 %.

In future work we can combine the two or more feature extraction methods to check whether it enhances result or not. We also collect the real time tweets to check the live streaming.

REFERENCES

[1] D.Chaffey, Global Social Media Research Summary 2016. URL ⟨http://www.smartinsights.com/Social-media-marketing/social-media-strategy/new-global-social-media-research/⟩.

[2] W.Tan, M.B.Blake, I.saleh, S.Dustdar, Social-network-sourcedbigdataana-lytics, InternetComput.17(5)(2013)62–69.

[3] S.K. Bharti B. Vachha , R.K. Pradhan , K.S. Babu , S.K. Jena "Sarcastic sentiment detection in tweets Streamed in real time: a big data approach", Elsevier 12 July 2016.

[4] Aditya Joshi, Vinita Sharma, Pushpak Bhattacharyya "Harnessing Context Incongruity for Sarcasm Detection" Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 757–762,Beijing, China, July 26-31, 2015.C 2015 Association for Computational Linguistic.

[5] Toma Ptacek Ivan Habernal and Jun Hong "Sarcasm Detection on Czech and English Twitter", Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 213–223, Dublin, Ireland, August 23-29 2014.

[6] R. Gonzalez-Ibanez, S. Muresan, and N. Wacholder. 2011. "Identifying Sarcasm in Twitter: A Closer Look".In Proceedings of the 49th Annual Meeting of Association for Computational Linguistics.

[7] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang, "Sarcasm as contrast between a positive sentiment and negative situation", in Proc. Conf. Empirical Methods Natural Lang. Process, Oct.2013,pp.704_714.

[8] Tony Veale and Yanfen Hao. 2010. Detecting Ironic Intent in Creative Comparisons. In ECAI, Vol. 215.765–770.

[9] A. Rajadesingan, R. Zafarani, and H. Liu, ``Sarcasm detection on Twitter A behavioral modeling approach," in Proc. 18th ACM Int. Conf. Web Search Data Mining, Feb. 2015, pp.79_106.

[10] M. Bouazizi, T. Ohtsuki, "Pattern-Based Approach for Sarcasm Detection on Twitter" VOLUME 4, 10.1109/ACCESS.2016.2594194.

[11] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Prateek Vij"A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks".Nanyang Technologica University 50 Nanyang Ave, Singapore 639798.

[12] B. Pang, L. Lee, S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," In Proceedings of the Conference on Empirical Methods in Natural Language Processing, July 2002, pp. 79-86.

[13] Kang Hanhoon, YooSeongJoon, Han Dongil, "Senti-lexicon and improved Naive Bayes algorithms for sentiment analysis of restaurant reviews", Expert SystAppl 2012, 39:6000 10.

[14] Y. Qiu, G. Yang, and Z. Tan, "Chinese text classification based on extended nave bayes model with weighed positive features," in First International Conference on Pervasive Computing, Signal Processing and Applications, 2010, pp. 243-246.

[15] Pooja Deshmukh, Sarika Solanke." Review Paper: Sarcasm Detection and Observing User Behavioral" Journal : International Journal of Computer Applications (0975 – 8887) Volume166–No.9,May2017.