Improvement Sarcasm Analysis using NLP and Corpus based Approach

Manoj Y. Manohar
Department of Computer Science and Engineering
Government Engineering College, Aurangabad
Maharashtra, India
E-mail: manoj.mym@gmail.com

Prof. Pallavi Kulkarni
Department of Computer Science and Engineering
Government Engineering College, Aurangabad
Maharashtra, India
E-mail: pallavi.k11@gmail.com

***Abstract*—Sarcasm is most important part of social network and microblogging website. Millions of people use it while using the social sites or twitter websites. Most of the sarcasm occurrence on the twitter. Most of the people express their thinking through twitter regarding any specific subject. It is the best way to convey the message to the any end user. Hence, finding the sarcastic statements is very useful in day-to-day life to improve the sentiment analysis from the sarcastic data from the twitter or any social websites. Sentiment Analysis indicates the expression of the user towards a specific topic. In this paper, we propose a NLP and CORPUS based approach to detect sarcasm on Twitter. In this we are comparing the data with the ontology based emotion detection and classify the tweets as a Sarcastic or non-Sarcastic. In particular, we emphasize the importance NLP and CORPUS based for the detection of sarcastic statements.**

Keywords- NLP; Corpus based approach; Emotion; Ontology

I. INTRODUCTION

In today's world, Twitter is one of the highest destinations for millions of people to describe their opinion, thoughts and to put their suggestions about real time occasions. And now day by day the use of twitter is increasing and that's why data also increasing which is called big data. Now from present days , according to official websites twitter has more than 288 million active users and more than 500 million tweets are sent every day. So many companies are attached to this data for knowing the opinion of the people toward movies[2], social work, political work[1] and product.

As many of informal language going to be used in twitter also there are too many limitations in terms of characters, feature based, understanding the opinion of users and performing analysis is quite difficult. Also existence of sarcasm in twitter create the task more difficult; we know the sarcasm is the way to make a joke of someone, to criticizing or humorous someone. Liebrecht et al. [3] discussed how sarcasm can be a polarity- switcher, and Maynard and Greenwood [4] proposed a set of rules to decide on the polarity of the tweet (i.e., whether it is positive or negative) when sarcasm is detected.

The online Oxford dictionary2 defines sarcasm as *``the use of irony to make or convey contempt''*. Collins dictionary3 defines it as *``mocking, contemptuous, or ironic language intended to convey scorn or insult''*. However, sarcasm is a deeper concept, highly related to the language and to the common knowledge. Although different from one another, sarcasm and irony have been studied as two close and much correlated concepts [4]_[6] or even as the same one [7]_[9]. The Free Dictionary4 also defines sarcasm as a form of irony that is intended to express contempt. Since most of the focus on sarcasm is to enhance and refine the existing automatic sentiment analysis systems, we also use the two terms synonymously.

As many people use twitter for specially use of conveying expression in the form of jokes, seriousness, humorous, criticized ,indicate their views on topics. So many of them use sarcastic statement as a tweet , means whatever they write , conclude the different meaning of that one. And such sarcastic tweets specially occurs on social websites such as twitters. Thus Maynard and Greenwood [4] said sentiment analysis should be increased sarcasm would find inside the real sarcastic statement. Hence, the necessity for an potential route to detect sarcasm arises.

In this work, we introduced an accomplished way to know tweets are sarcastic tweet or non-sarcastic tweet. Even though, no need an already-built user knowledge base as in the work of Rajadesingan et al. [5], our method recognize the different types of sarcasm and detect the sarcastic tweets and non-sarcastic tweets regardless of their owners or their temporal context,

Here upon, the main subscription of this paper are as follows:

- To detect the intention behind use of sarcastic statement into the tweets by people.
- Also to propose an potential way to detect tweets as sarcastic or non-sarcastic, and by using this information we can study for how can we increased accuracy of sentiment analysis.
- And by use of various feature how should we increased the accuracy of the method.

II. MOTIVATION

As we know that, in present days twitter or social websites are going to be widely used for conveying their thoughts towards any specific topic. So, millions of people are using the twitter for put their thinking and sometimes what they write , the meaning comes the different than writing statement.

978-1-5386-2745-7/17/$31.00 ©2017 IEEE

So, it is very important to know the meaning behind every tweet of the millions of people what about they tweet. And our proposed approach is to identification of the tweet is sarcastic or non-sarcastic and definitely it will help to renovate sentiment analysis of the millions of people about their tweets. In Twitter, sarcastic texts are very common. *``All your products are incredibly amazing!!!''* might be considered as a compliment or not. So whatever the user are saying through tweet, sometimes he means to say totally opposite to the statement. Thus, it might be inevitable to search a route to automatically explore the sarcastic tweets.

Rajadesingan et al. [5] have also mentioning the limit of their "A behavioral modeling approach" which does the sentiment analysis. Also they mention why sarcasm is difficult to find by humans and how it looks to be complicated. And that's why there is necessity to find out the sarcastic and non-sarcastic messages into the twitter.

Unless, many objections comes and create the task very complex. Joshi et al. [6] have mentioned 3 main difficulties which are i) the recognisance of common words, ii) the intent to redicule , and iii) the speaker-listener (or reader in the case of written text) context.

In this work, we present Natural Language Processing and NLP application on tweets to fetch the action word and Once we get the action words from tweets, then we will compare them with a corpus of sarcasm data using semantic matching and graph based matching and This will give a score of sarcasm for the given tweet and Using this score we will be able to detect the level of sarcasm in the given tweet.

III. RELATED WORK

More attention provided in the last some years to twitter sentiment analysis by investigator or researchers and also too many papers are going to be addressing to the classification of the tweets as sarcastic or non-sarcastic. However, the nature of the classification and the features used vary depending on the aim. Sriram et al. [18] used non-context-related features such as the presence of slangs, time-event phrases, opinioned words, and the Twitter user information to classify tweets into a predefined set of generic classes including events, opinions, deals, and private messages. Accra et al. [15] proposed a method to identify the emotional pattern and the word pattern in Twitter data to determine the changes in public opinion over the time. They implemented a dynamic scoring function based on Jaccard's similarity [5] of two successive intervals of words and used it to identify the news that led to breakpoints in public opinion. However, most of the works focused on the content of tweets and were conducted to classify tweets based on the sentiment polarity of the users towards specific topics. A variety of features was proposed. Not only they include the frequency and presence of unigrams, bigrams, adjectives, etc. , but they also include non-textual features such as emoticons [18] (i.e., facial expressions such as smile or frown that are formed by typing a sequence of keyboard symbols, and that are usually used to convey the writer's sentiment, emotion or intended tone) and Dong et al. [19] proposed a target-dependent classification framework which learns to propagate the sentiments of words towards the target depending on context and syntactic structure. Sarcasm, on the other hand, and irony in general have been used by people in their daily conversations for a long time. In this context, researchers have recently been interested in sarcasm, trying to and ways to automatically detect it when it is present in a statement. Although some studies such as [4] highlighted that, unlike irony, sarcasm *``is not a discrete logical or linguistic phenomenon''*, many works have been proposed and present high accuracy and precision. Burfoot and Baldwin [13] introduced the task of filtering satirical news articles from true newswire documents. They introduced a set of features including the use of profanity and slangs and what they qualified of *``semantic validity''*; and used Support Vector Machine (SVM) classifier to recognize satire articles. Campbell and Katz [14] studied the contextual components utilized to convey sarcastic verbal irony and proposed that sarcasm requires the presence of four entities: allusion to failed expectation, pragmatic insincerity, negative tension and presence of a victim, as well as stylistic components. Nevertheless, other works have been proposed to represent sarcasm. Some of these representations are given in [6] as follows:

Wilson also mentioned that, sarcasm arises when there is situational difference between text and theme.

Ivanko and Pexman expressed that sarcasm requires a 6-tuple consisting of a speaker, a listener, a context, an utterance, a literal proposition and intended proposition.

Giora manifested that sarcasm is a form of negation in which an explicit negation marker is lacking. This implies that the sarcasm is namely a polarity-shifter.

As for the task of detection itself, several goals were defined. Tepperman et al. [15] studied the occurrence of the expression *``yeah right!''*, and whether it appears in a sarcastic context or not. They proposed an approach to automatically detect sarcasm present in spoken dialogues, using prosodic, spectral and contextual cues. However, this represents the main shortcoming for their approach: absence of such components makes it impossible to detect sarcasm. In other words, although the approach itself is very effective in detecting when a specific expression is sarcastic, this approach is unable to detect all types of sarcasm that might occur.

Ghosh et al. [16] mentioned the words have a literal spirit and here upon by finding the spirit of word, sarcasm can be detected.

Maynard and Greenwood [4] remained on hash tags and there are millions of people of twitter users introduced an hash tags in their tweets to know the sarcasm in the tweets which are put up by the people into the tweets. Also they mentioned about how recognition of sarcasm can increased the sentiment analysis of tweets.

Riloff et al. [17] proposed a "Sarcasm as contrast between a positive sentiment and negative situation". This method detect a sarcasm, where a positive sentiment contrasts with a negative situation. Also using single seed word "love" i.e.

978-1-5386-2745-7/17/$31.00 ©2017 IEEE

bootstrapping algorithm and a drift of too many tweets i.e. sarcastic to automatically detect and learn expressions showing positive sentiment and phrases citing negative situations. Their approach shows some potentials. However, most of the sarcastic tweets in Twitter do not fall in the aforementioned category of sarcasm. In addition, the approach relies on the existence of the all possible ``*negative situations*'' on the training set, which makes it less efficient when dealing with new tweets.

Rajadesingan et al. [5] went deeper and dealt with the psychology behind sarcasm. They proposed ``Sarcasm detection on Twitter: A behavioral modeling approach'' for sarcasm detection on twitter. It has finding the different manner of sarcasm and their expression on twitter and explained the requirement of historical information collected from the past tweets. Even if, it proven to be very accomplished, the method is less performant, when if there is no previous data about the user. Most of the characteristics extracted depend on data collected from previous tweets to judge. Also the real time flow of tweets, where aimlessly users are posting tweets, so it is hard to run the approach, the size of the knowledge-base grows very fast, and the training should be redone each time based on the new tweets collected (i.e., since the previous tweet has the highest impact on the current one, the new tweet should be taken into consideration for the next iteration).

IV. PROPOSED WORK

In this proposed method, intention is to classify the tweets as a sarcastic or non-Sarcastic. The tweets are collected by querying Twitters Streaming API. These tweets are collected and stored into the tweet folder. Following are the necessary steps for work:-

A. Data

Data are the tweets as a database receives using the streaming API available by twitter to get access to real time tweets for performing operations.

B. Data Pre-processing

Twitter data can contain various texts or(using @<user>) tags called hashtags, for example #understand, #beautiful. Hence, it is important to process the data before applying to feature extraction. Data preprocessing includes cleaning, Instance selection, normalization, transformation, feature extraction and selection, etc.

C. Algorithm

Step 1: Tweet fetching from Twitter
Step 2: NLP application on tweets to fetch the action words
Step 3: Once we get the action words from tweets, then we will compare them with a corpus of sarcasm data using semantic matching and graph based matching
Step 4: This will give a score of sarcasm for the given tweet
Step 5: Using this score we will be able to detect the level of sarcasm in the given tweet

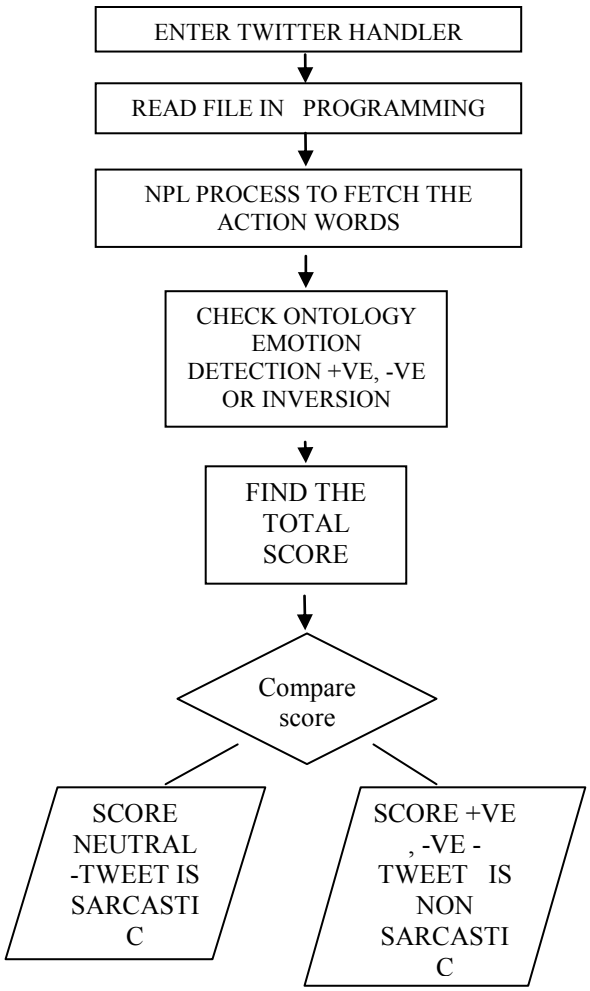
Using the algorithm and the emotion word ontology we process two types of input data:

1) Text: The text data is directly given as an input. The input for text can be taken from newspaper articles, documents, files etc.

2) Tweets: The tweets are extracted from the twitter using a twitter handler and the tweets are given as input. We use the Alchemy API which is used to provide the web services by communicating with twitter to retrieve tweets.

D. System model

Model used to finding the Tweet is sarcastic or Tweet is non-Sarcastic. We can access the Tweets through Twitter handler and millions of people use twitter to express their feelings regarding specific topic. There are various purposes behind the use of twitter. So, here using the NLP Classifier to making the tagging. Various features extracted while using NLP on tweets to finding the tweet is sarcastic or non-Sarcastic.



978-1-5386-2745-7/17/$31.00 ©2017 IEEE

Figure1: System model for detection of sarcasm on Twitter

a) NLP Classifier

The classifier for natural language processing is a tool that takes in the data-items and categorizes them into one of the k-classes. In natural language processing (NLP) we perform various tasks such as tokenization, part of speech tagging (PoS), lemmatization, etc.

b) Tokenization

Tokenization is the procedure of split-up given sentences of tweets into pieces called tokens. It occurs normally at word level. Also token called as sequence of characters that group together to form a useful semantic unit which can later be used for processing. Tokenization split the flow of text into words, phrases, or other meaningful components referred to as tokens in lexical analysis. List of token is an output of a tokenization becomes input for forward analysis. Some of the heuristics used by tokenizer are as follows (i) the whitespace characters such as a space or line break are used to separate the tokens, (ii) tokens can be composed of contiguous strings of alphabetic characters or numbers or alphanumeric character, (iii) tokens may or may not include punctuations and whitespaces.

c) Part-Of-Speech (POS) Tagging

Part-Of-Speech (POS) Tagging play a very important role into the text processing. POS tagging is also called by word-category disambiguation. POS play an very important role like reading the text and commend parts of speech such as verb, noun, conjunction, adjective, interjection, adverb, etc. to each token based on definition and relationship adjacent and related words in the sentence. Many experiments in computational science require POS tagging, for example, nouns can be distinguished as the plural, possessive, and singular forms. It makes uses NN for singular common nouns, NNS for plural common nouns and NP for singular proper nouns. In this work, tagging the text with standard PoS tagger in java applications. To make such task tagger has to load the "trained" file that contains important information for the tagger to tag string. This "trained" file is called a model. There are many trained models provided by Stanford NLP group for different language.

d) Corpus based method

The input dataset is first stored into the memory when the user gives an input then the tweets are fetched for each tweet we compare it with the stored dataset and find out the required sentiment from it
based on the sentiment aggregation we check if the tweets are sarcastic or not.

e) Performance Measures

Accuracy: high accuracy is achieved when the results returned by the algorithm are significantly more relevant than irrelevant ones.

$$\text{Accuracy} = \frac{\text{No. of sarcastic tweets}}{\text{Total no. of Tweets}}$$

V. EXPERIMENTAL RESULT

Overall performance of the proposed approach as shown in the table below. It indicate (+) sign as a sarcastic Tweet manually and (-) sign indicate the non-Sarcastic Tweet after running. In this, we are using real time database, the tweets are collected by querying Twitters Streaming API. The java language is used to develop the system along with the NLP model.

In this table taken the 10 tweet with twitter name and every twitter name mentioning the number of tweets to be contained and every tweet have execution time. Fig. 2 indicate the relation between tweet and delay.

Table 1. Performance of the proposed approach

Twitter Id	Twitter Name	No. Of Tweets	Effect Observed	Effect Obtained	Delay
1	cyberDomain	20	+	+	877
2	Goldberg	25	+	+	444
3	iForex_com	20	-	+	429
4	manoj	40	-	-	411
5	RafaelNadal	20	+	-	409
6	realDonaldTrump	20	+	+	426
7	ReutersTV	20	+	-	688
8	SrBachchan	24	-	-	567
9	The_Rock009	9	+	+	379
10	Vishalnell	1	-	-	389

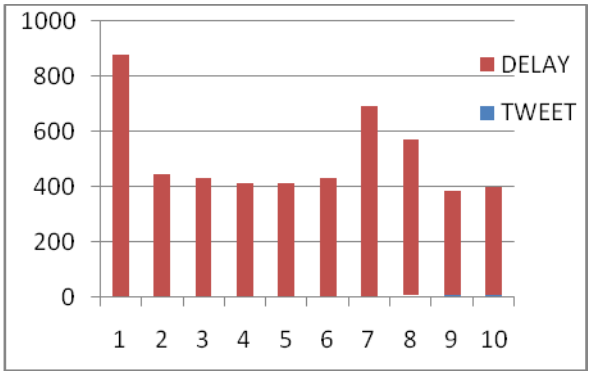


Fig 2:- Relation between Tweet and Delay

VI. CONCLUSION

In this project, we proposed new techniques to detect the sarcasm on the real database of the Twitter. The propose techniques making use of the action words by using Part-of-Speech tags. The technique is showing the good result on real time database of Twitter. Also we are using efficient techniques such as ontology based emotion detection to improve the sarcasm on the Twitter. Objective of the propose work is to fetch tweets for a particular person and check out sentiments of each tweet to evaluate sarcasm of each person. Also there are various significances of the sarcasm detection such as, personality analysis, twitter profile analysis. And future work to empower the performance of sentiment analysis and opinion mining.

REFERENCES

[1] J. M. Soler, F. Cuartero, and M. Roblizo, ``Twitter as a tool for predicting elections results," in *Proc. IEEE/ACM ASONAM*, Aug. 2012, pp. 1194_1200.

[2] U. R. Hodeghatta, ``Sentiment analysis of Hollywood movies on Twitter," in *Proc. IEEE/ACM ASONAM*, Aug. 2013, pp. 1401_1404.

[3] C. C. Liebrecht, F. A. Kunneman, and A. P. J. van den Bosh, ``The perfect solution for detecting sarcasm in tweets #not," in *Proc.WASSA*, Jun. 2013, pp. 29_37.

[4] D. Maynard and M. A. Greenwood, ``Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis," in *Proc. 9th Int. Conf. Lang. Resour. Eval.*, May 2014, pp. 4238_4243.

[5] 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.

[6] A. Joshi, P. Bhattacharyya, and M. J. Carman. (Feb. 2016). ``Automatic sarcasm detection: A survey." [Online]. Available: https://arxiv.org/abs/1602.03426

[7] B. Pang, L. Lillian, and V. Shivakumar, ``Thumbs up?: Sentiment classification using machine learning techniques," in *Proc. ACL Conf. Empirical Methods Natural Lang. Process.*, vol. 10. Jul. 2002, pp. 79_86.

[8] Rubo Zhang, Ying Wang, Jing Wang, (2008), "Research on Ontology Matching Approach in Semantic Web", International Conference on Internet Computing in Science and Engineering, IEEE 978-0-7695-3112-0/08.

[9] M. Bouazizi and T. Ohtsuki, "Sarcasm detection in Twitter," IEEE GLOBECOM 2015, to be published.

[10] M. Bouazizi and T. Ohtsuki, "OPINION MINING in Twitter," IEEE/ACM 2015, to be published.

[11] A. Joshi, V. Sharma, and P. Bhattacharyya, ``Harnessing context incongruity for sarcasm detection," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics, Int. Joint Conf. Natural Lang. Process. (ACL-IJCNLP)*, vol. 2. Jul. 2015, pp. 757_762.

[12] M. W. Berry, Ed., *Survey of Text Mining: Clustering, Classification, and Retrieval*. New York, NY, USA: Springer-Verlag, 2004.

[13] C. Burfoot and T. Baldwin, ``Automatic satire detection: Are you having a laugh?" in *Proc. ACL-IJCNLP*, Aug. 2009, pp. 161_164.

[14] J. D. Campbell and A. N. Katz, ``Are there necessary conditions for inducing a sense of sarcastic irony?" *Discourse Process.*, vol. 49, no. 6, pp. 459_480, Aug. 2012.

[15] J. Tepperman, D. Traum, and S. S. Narayanan, ```Yeah right': Sarcasm recognition for spoken dialogue systems," in *Proc. InterSpeech*, Sep. 2006, pp. 1838_1841.

[16] D. Ghosh, W. Guo, and S. Muresan, ``Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words," in *Proc. EMNLP*, Sep. 2015, pp. 1003_1012.

[17] 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.

[18] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, ``Short text classification in twitter to improve information ltering," in *Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2010, pp. 841_842.

[19] L. Dong, F.Wei, C. Tan, D. Tang, M. Zhou, and K. Xu, ``Adaptive recursive neural network for target-dependent Twitter sentiment classification," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, vol. 2. Jun. 2014, pp. 49_54.