An Emoticon based Sarcastic Sentiment Analysis on Live Streaming Data of Twitter

Dr. Smita Rani Sahu[1] Gadepalli Anuradha Mrudula[2]
[1,2]Department of Computer Science & Engineering
[1,2]Sri Venkateswara College of Engineering, Andhra Pradesh, India

Abstract— Sentiment Analysis (SA) is a methodology in the data mining field. It plays a crucial role in the opinion mining, statistical experiments especially in the field of text mining. Generally an analysis can be made depending on the requirement given for the user. Coming to the social media, it is becoming a vast space for the people to share their views and ideas. Intentionally they make comments like reviewing a particular topic regarding its positive and negative thoughts. This can be made by Sentiment Analysis which is used in determining the opinions of the commentators. In this paper, an analysis is done on the sarcastic opinions or reviews based on the] emoticons and short texts, which are now very familiar in posted in a social media like Twitter. Twitter will be having many tweets regarding various cases like movies, restaurants, particular person, etc. A text in the review may not give the full result regarding an opinion. So, emoticon can give maximum accuracy in determining the sarcastic opinions of the users. The main intention of the paper is to analyze the tweets called as reviews of the specific topic dynamically and calculating the review rate using the combination of short texts and emojis.

Key words: Sentiment Analysis, Sarcasm, Data Mining, Lexicon, Corpus Based, Emoticons, Tweets, Social Media, Short Texts

I. INTRODUCTION

Twitter has provided more space for prediction of consumer brands, movie reviews, democratic electoral events, stock market, and popularity of celebrities. It is very difficult to analyze this bulk of user generated data manually. We introduce approach for automatically classifying the sentiment of [12] Twitter messages. These messages are classified as either positive or negative with respect to a review. This is useful for every person who want to explore the sentiment of various products before purchase, or companies that want to monitor the public sentiment of their brands. So there is a need to develop an intelligent system which automatically mines such huge content of attitudes, opinions, and emotions from text, and speech and database sources through Natural Language Processing (NLP).

In the existing system sentiments are analyzed based on text. Earlier, tweets are like example say "this movie is boring". But now, the tweets are like "GM! This movie not bad - -". In this review we see combination of text, short text and emoticon. So it is very difficult to analyze exact opinion of the user using only text in the real time which contains huge data. It is hard to calculate if the comment includes [8] emoticons and short texts. The proposed work is able to collect information from Twitter for finding the sentiments and we analyze sentiments on text as well as emoticons. These emoticons give the actual feeling of the commenter's as well as the accuracy in finding the sentiment will be increased. Since the process makes use of emoticons as noisy labels, it is crucial to discuss the role they play in classification of sentiments.

II. RELATED WORK

A lot of work has been done in the field of [11] opinion mining or sentiment analysis for well over a decade now. Different techniques are used to classify the text according to polarity. Most of these techniques can be classified under two categories: Machine Learning and Lexicon Based. Some of the techniques used for different purposes in sentiment analysis algorithms is described below in Table 1.

Algorithm	Dataset	Scope
[15] Lexicon - based, semantic	Experienceproject.com	Stories
[16] Lexicon - based, semantic	IMDB	Movie Reviews
[17] Statistical (MM), semantic	Amazon	Product Reviews
[18] Statistical	Amazon	Book Reviews
[19] Corpus-Based	Live journal	Blog
[20] Corpus-Based	Twitter	Tweet Reviews
[21] Lexicon - Based, SVM	Dutch wordnet	Lexicons

Table 1: Various Algorithms used for Different Analysis

Machine learning techniques first trains the algorithm with some particular inputs with known outputs so that later it can work with new unknown data. That is, a machine learning algorithm needs to be trained first for both supervised and unsupervised learning tasks. [1] Naïve bayes is one of the most popular method in text classification. It is considered as one of the most simple and efficient approaches in NLP. It works by calculating the probability of an element being in a category. First the prior probability is calculated which is afterwards multiplied with the likelihood to calculate the final probability. The method assumes every word in the sentence to be independent which makes it easier to implement but less accurate. [2] The sentiment classification was done using discriminative classifier. This approach is based on structural risk minimization in which support vectors are used to classify the training data sets into two different classes based on a predefined criteria. [3] Shi used centroid classifier which assign a centroid vector to different training classes and then uses this vector to calculate the similarity values of each element with a class. If the similarity value exceeds a defined threshold then the element is assigned to that class (polarity in this case). Songho also used K-Nearest Neighbor (KNN) approach which finds 'K' nearest neighbors of a text document among the documents in the training data set. Then classification is performed on the basis of similarity score of a class with respect to a neighbor.

A text phrase was used as a query for a search engine and the results were used to classify the text. Twitter data was used for sentiment analysis by Negation word can reverse the polarity of any sentence. Lexicon Based techniques work on an assumption that the collective polarity of a document or sentence is the sum of polarities of the individual words or phrases. [4] Kamps used a simple technique based on lexical relations to perform classification of text. [5] Andrea used word net to classify the text using an assumption that words with similar polarity have similar orientation. [6] T. Wilson used an algorithm based on pos (part of speech) patter. It was observed that most of these existing techniques doesn't scale to big data sets efficiently. While various machine learning methodologies exhibits better accuracy than lexicon based techniques, they take more time in training the algorithm and hence are not suitable for big data sets.

A. Limitations

An individual's sentiment towards a brand or product may be influenced by one or more indirect causes; someone might have a bad day and tweet negative remark about something they otherwise had a pretty neutral opinion about. That time it is very difficult to analyze the opinion of user.

The opinion of user will be changes over time according to a person's mood. So, it is difficult to analyze the opinion of user. The limitations of the existing system are the indeed outcome of our paper.

III. PROPOSED WORK

The main objective of the project is that it extracts the tweets and analyses the sentiments in both text as well as emoticons. Due to this we can find the exact opinion of the user. Companies mostly benefit from sentiment analysis today. The key points show that businesses can solely track positive and negative reviews of their brands. With that being said, it also helps them measure their overall performance, especially on their online present consumers can use sentiment analysis to research products or services before making a purchase. Marketers can use this to research public opinion of their company and products, or to analyze customer satisfaction. Organizations can also use this to gather critical feedback about problems in newly released products. On the other hand, certain individuals can also get a lot from sentiment analysis, whether they are making a brand for themselves or just having that drive to know anything that regards to them. An ordinary person, say a fanatic or blogger, can also benefit from sentiment analysis. Usually the Lexicon based approach for [9] Sentiment Analysis is given in two kinds like dictionary based and Corpus based approaches. The [13] Dictionary based approach is the base for Corpus based approach. In this all the datasets are derived using dictionaries where common sets like stop words, positive words, negative words, etcetera are used. The [7] Sarcastic Sentiment analysis is done by the [14] Corpus based algorithm considering the tweets with list of likes and dislikes emoticons and short texts including dictionaries as an input parameter for the particular user. Corpus based suggests data-driven approach where you will have access not only to sentiment labels, but to a context which you can use to your advantage in an ML algorithm. It can certainly be a rule-based approach with NLP parsing too,

or even a combination. Corpus also carries some domain specificity that can inform your algorithm of sentiment label variety for a word depending on its context / domain. The sample emoticons list is displayed in Table 2.

– Happiness	:-D, =D, xD, (^ ^)
– Sadness	:-(, =(
– Crying	:'(, ='(, (; ;)
– Boredom	- -, -.-, (> <)
– Love	<3, (L)
– Embarrassment	:-$, =$, >///<

Table 2: Sample list of Emoticons

While testing sarcasm in tweets, one needs to calculate the sentiment score of the tweet. Then, extract the triplet (subject, verb and object) of that tweet. Here we more concentrated on stop words which are all included with triplet. Short texts (Table 3) are also verified to get the more accurate opinion of the user. The short texts are taken as three varieties like positive, negative and neutral. These short texts are counted with the emoticons to calculate the text or document as the positive or negative sentiment expressed by the commentator.

Positive	Negative	Neutral
GM (Good Morning)	BH (Block Head)	QOTD (Quote of the day)
TQ, TY (Thank You)	YKYAT (You know you are addicted to)	KK (ok, got it)
HAND (Have a nice day)	SRS (Serious)	IC (I see)
LOL (Laugh out Loud)	NCT (Nobody cares Though)	RE (Reply in regards)

Table 3: Sample List of Short Texts

If the tweet is positive and the subject is not a pronoun check the subject value in the positive list. If the subject value is found in the likes list i.e., positive dictionary the tweet is not sarcastic. If it is found in the dislikes list i.e., negative dictionary, the tweet is sarcastic. Similarly, if the subject value is a pronoun and the tweet is positive the object value checks the likes list. If it is found the tweet is not sarcastic. If it is found in the dislikes list the tweet is sarcastic. In a similar fashion, one identifies sarcasm for negative reviews also. In this paper we used (Twitter 4j) API to connect to the cloud for retrieving opinions. The complete Java in IDE NetBeans environment gives a smooth process of connecting to the API. The coming section describes the algorithm of the approach which is used in our analysis.

IV. PRACTICAL APPROACH

The Corpus Based Approach Algorithm is given below. And the result produces a graph which represents the sarcastic analysis of the tweets or reviews. The graph is shown in the below figure Fig.1.
1) Data: dataset:= Corpus of Tweets, LL, DLL
2) Result: Classification:= Sarcastic or not Sarcastic
3) Notation: N: Words, T: Tweets, C: Corpus
4) PSTF: Positive_Sentiment_Tag_File,

5) NSTF: Negative_Sentiment_Tag_File,
6) TWT: Tweet_wise_tag,
7) FT: First_tag,
8) PSEF: Positive_Sentiment_Emoticon_ File,
9) NSEF: Negative_Sentiment_Emoticon_File,
10) LL: Likes_list, DLL: Dislikes_list.
− Initialization
$PSTF=NSTF=PSEF=NSEF=\{\varphi\}$
while T in C do

       SC= find_sentiment_score(T)

       k=find_Postag(T)

       x=find_parse(T)

       t=find_subset(k)

       FT=find_first_tag(t)

       if($FT=pronoun$) then

            O= find_object(x)

       end

       else

            S= find_object(x)

       end

       if(SC>0.0) && (O in LL) then

            Tweet is Sarcastic

       end

            Discard the tweet

       end

       if(SC>0.0) && (O in DLL) then

            Tweet is not Sarcastic

       end

            Discard the tweet

       end

       if(SC<0.0) && (O in LL) then

            Tweet is Sarcastic

       end

            Discard the tweet

       end

       if(SC<0.0) && (O in DLL) then

            Tweet is not Sarcastic

       end

            Discard the tweet

       end

       if(SC>0.0) && (S in LL) then

            Tweet is Sarcastic

       end

            Discard the tweet

       end

       if(SC>0.0) && (S in DLL) then

            Tweet is not Sarcastic

       end

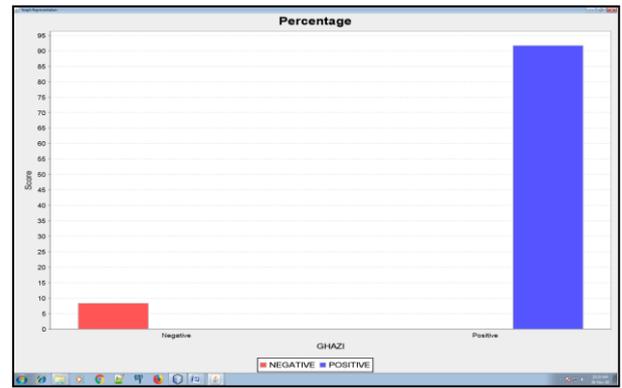   Discard the tweet

       end

end

Fig. 1: Graph Showing Result of Sarcastic Sentiment Analysis

V. CONCLUSION & FUTURE ENHANCEMENT

In this paper we discussed about the accuracy of the user's opinion based on the Sarcasm present in the tweets of the Twitter. Here we showed that emoticons and short texts are used for calculations. But there is a lot of things to be analyzed more deeply to calculate the accuracy like scaling. Now a day's people are in advance that they are giving reviews based on scaling for example. If the topic is good, comments as 5, poor -1, very good-6, etc., according to the domain they need (like twitter they scale their comments from 1 to 10 assume). In future this must be useful for every individual in a way that user can expect correct opinion about the specific topic in the required domain.

REFERENCES

[1] Zhaoxia WANG, Victor Joo Chuan TONG, David CHAN, 2014 IEEE 6th International Conference on Cloud Computing Technology and Science
[2] V.K. Singh, R. Piryani, A. Uddin, Sentiment Analysis of Movie Reviews, 978-1-4673-5090-7/13/$31.00 ©2013 IEEE
[3] H. X. Shi, G. D. Zhou, and P. D. Qian, "An attribute-based sentiment analysis system," Information Technology Journal, Vo1.9, No.8, pp.1607-1614,2010
[4] Kamps, M., Marx, R.J. Mokken, and Rijke, M.D., "Using wordnet to measure semantic orientation of adjectives", proceeding in National Institute for (2004) 1115– 1118, 2004
[5] Andreas, H., Gerhard, P., and Andreas, N., "A Brief Survey of Text Mining", LDV Forum - GLDV Journal for Computational Linguistics and Language Technology, 2005
[6] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis," Association for Computational linguistics, vol. 35, no. 3, 2009.
[7] Piyoros Tungthamthiti, Kiyoaki Shirai, Masnizah Mohd, Recognition of Sarcasm in Tweets Based on Concept Level Sentiment Analysis and Supervised Learning Approaches, 28th Pacific Asia Conference on Language, Information and Computation pages 404–413
[8] Wiesław Wolny, Sentiment Analysis of Twitter data using Emoticons and Emoji ideograms, ISSN 2083-8611

[9] Y.L. Zhang, Y. Dang, H.C. Chen, "A Lexicon-Enhanced Method forSentiment Classification: An Experiment on online Product Reviews," IEEE Intelligent Systems 2010, Jul. 2010, pp. 46-53

[10] Y. Yu, W. Duan, and Q. Cao, "The impact of social and conventional media on firm equity value: A sentiment analysis approach," Decision Support Systems, vol. 55, no. 4, pp. 919-926, Nov. 2013

[11] B. L. Xiaowen Ding and P. S. Yu. A holistic lexicon-based approach to opinion mining. In Proceedings of the international conference on Web Search and web Data Mining, pages 231–240. ACM, 2008.

[12] Davidov D, Tsur O, Rappoport A. Enhanced sentiment learning using twitter hashtags and smileys[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010: 241-249

[13] Fatehjeet Kaur Chopra, Rekha Bhatia, Sentiment Analyzing by Dictionary based Approach, International Journal of Computer Applications (0975 – 8887) Volume 152 –No.5, October 2016

[14] 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, Volume 2, Issue 3, August 2016, Pages 108-121

[15] Neviarouskaya Alena, Prendinger Helmut, Ishizuka Mitsuru. Recognition of Affect,Judgment, and Appreciation in Text.In: Proceedings of the 23rd international conference on computational linguistics (Coling 2010), Beijing; 2010. p. 806 –14

[16] Heerschop B, Goossen F, Hogenboom A, Frasincar F,Kaymak U, de Jong F.Polarity Analysis of Texts using Discourse Structure. In: Presented at the 20th ACM Conference on Information and Knowledge Management (CIKM"11); 2011

[17] Zirn C, Niepert M, Stuckenschmidt H, Strube M. Fine-grained sentiment analysis with structural features. In: Presented at the 5th International Joint Conference on Natural Language Processing (IJCNLP"11); 2011

[18] Nan Hu, Indranil Bose, Noi Sian Koh, Ling Liu "Manipulation of online reviews: an analysis of ratings, readability, and sentiments" in Decis Support Syst, 52 (2012), pp. 674–684

[19] Fazel Keshtkar, Diana Inkpen "A bootstraping method for extracting paraphrases of emotion expressions from texts",Comput Intell, vol. 0 (2012).

[20] Alexander Pak, Patrick Paroubek, "Twitter as a Corpus for Sentimental Analysis and Opinion Mining" 81

[21] Alexandra Balahur, Jesús M. Hermida, Andrés Montoyo in "Detecting implicit expressions of emotion in text: a comparative analysis", Decis Support Syst, 53 (2012), pp.742–753