Public Sentiment Insights Analysis using Word Sense Disambiguation Application on Twitter Data during a Pandemic – Covid'19

K Anuratha[1], M Parvathy[2], R Priyadarshini[3], D Saisanthiya[4]

[1]Assistant Professor, Department of Information Technology, Sri Sai Ram Institute of Technology, Chennai, India
[2]Professor, Department of Computer Science and Engineering, Sethu Institute of Technology, Madurai, India
[3]Assistant Professor (Selection Grade), Department of Information Technology, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India
[4]Assistant Professor, Department of Information Technology, School of Computing, SRM Institute of Science and Technology, Chennai, India

ABSTRACT

The entire world is affected because of the global pandemic Covid-19 due to the virus belongs to the family of Coronavirus. As the spread of infection and mortality rate is rapid people have started developing assorted emotions about the crisis. It is more significant to administer the mental health and Psychological wellbeing of public during a crisis like this. As many of the people broadly use the social media like twitter for sharing their opinions and thoughts, our work utilizes the Covid specific Tweets posted by the Tweeple and analyse them to understand the sentiments exhibited regarding the situation. After the tweets are collected and the real sentiments behind them are discovered using the classifier model developed using the Machine Learning methods. The experimental results may be used by the respective authorities to take necessary initiatives for addressing the concerns that affect the wellbeing of the society and the economic wellbeing as well. As our word uses Lexical based sentiment analysis, it is important to remove the ambiguities of a word which is a main challenge of this technique on sentiment analysis. To improve the performance of the Sentiment Analysis we have used the lexical dictionaries Wordnet and SentiWordNet along with Word Sense Disambiguation (WSD) to detect and remove the ambiguities understanding the context of the term used in the tweets.

Key words: Sentiment Analysis, Word Sense Disambiguation (WSD), Opinion Mining, Twitter, Covid-19 and Coronavirus.

1. INTRODUCTION

Sentiment Analysis has drawn attention from researchers and has transformed the way in which the data are collected and opinions (that are conveyed directly / indirectly are extracted in social media) are analysed [1]. With the vast amount of data that is being produced every day, the data analysts have started mining the social media for useful information from the pool of data which will be used for interpreting the various opinions of users of social media platforms [2]. In this Era of Social Media, Twitter has become one among the widespread micro-blogging platforms and is used as the pool of information for performing Sentiment Analysis.

Sentiment Analysis focuses on finding the attitude of the writer on a specific topic under discussion or the content of a document as a whole. It is the application of Natural Language Processing (NLP) and computational linguistics to discover and extract the opinions or sentiments from the source information. The important task of Sentiment Analysis is to classify the texts n the given document, sentence or feature as positive, negative or neutral considering the polarity of the opinion of texts [3].

The techniques for Sentiment Classification are categorized as Machine Learning, Lexicon Based and Hybrid Approaches [4]. The Machine Learning technique uses linguistic features and Machine Learning Algorithms. The Lexicon approach depends on the sentiment lexicon, a set of sentiment terms that are already known. It has two methods further, dictionary based and corpus based that uses the semantic method to know the sentiment polarity. The combination of ML and Lexicon based approaches is the Hybrid approach. The measures of accuracy for Sentiment Analysis are Precision and Recall [5].In this work we have used lexicon-based approach that has the ambiguity issue as a major challenge that influences the accuracy of the Sentiment Analysis. The research on handling the ambiguity problems that affects the Sentiment Analysis accuracy is limited and still going on.

There are many words that refer to different senses in different context in all the major regional languages. Word Sense Disambiguation is the NLP process of extracting the appropriate sense of the ambiguous word in a given context using the predefined word senses set. In this work, we have introduced a Sentiment Classifier with Word Sense Disambiguation that uses the lexical sources WordNet and SentiWordNet for discovering sentiments from tweets on Covid-19. We have chosen the Twitter Platform as it has limited lengthen texts and relevance to content [6]. The tweets posted by the people can be further used by the respective authorities to gain insights on the mindset of public and to decide on the assistance policies needed [7].

Section 2 describes the Literature Review, Section 3 is about the methodology used and Section 4 discusses the details on Implementation and results. The work is concluded in Section 5.

2. LITERATURE REVIEW

To mine insights Dong et al [8] apply sentiment analysis in user reviews to suggest products and show the benefits of the work in Amazon products domain. Farhan Hassan Khan et al [9] proposed a Classification algorithm that incorporates improved form of emotion analysis using SentiWordNet analysis and an improvised classifier based on polarity. The result indicates that the algorithm overrides the earlier limitations and achieves greater accuracy. To understand the behaviour people and the societal culture during a crisis Hyo Jin Do et al [10] proposed an opinion analysis during the pandemic outbreak in Korea. To monitor the transmission of epidemic in people twitter data was used by Vasileios Lampos et al [11] . Felipe Taliar [12] collected posts from facebook and classified the reactions into positive, negative and neutral.

3. METHODOLOGY

The work mainly focuses on Extracting Tweets and Sentiment Analysis. To extract the true sense of words, semantics is incorporated in Sentiment Analysis using a dictionary of sentiments SentiWordNet with Word Sense Disambiguation. The System is then used for measuring the sentiments in Covid-19 related tweets. The system has three parts.

1. Acquiring Data
2. Pre-processing
3. Classifying Sentiments
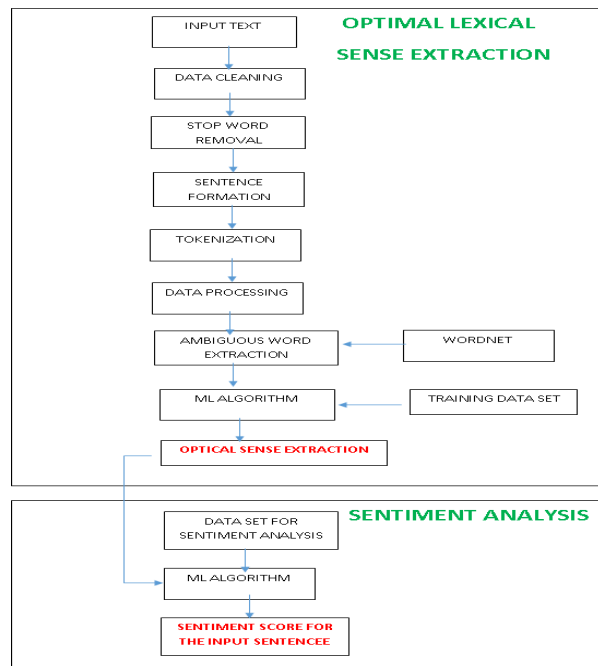
Figure 1: Block Diagram of Proposed Work

The block diagram for the work to classify sentiments is shown in Figure 1.The classifier is developed to discover sentiments of people during a pandemic from extracted tweets related to Covid-19.

1. Acquiring Data

We have used Twitter streaming API, which allows the access to freely accessible data on Twitter to obtain the tweets continuously [13].The next part, Pre-processing takes these tweets as input. Dataset that includes COVID-19 related tweets during First four months (JAN – APRIL) of 2020 is taken from TweetBinder, for experimentation.

2. Pre-Processing

In this step, the URLs present in the Tweets, private usernames start with @, all the hash tags start with # and all the special characters, if any are removed. The pre-processed tweets are the input for the next phase which is the Sentiment Classification.

3. Classifying Sentiments

We have adapted the lexicon based approach for classifying sentiments with the concept of Word Sense Disambiguation using WordNet and SentiWordNet lexical sources [14,15]. The words of English Language are grouped into set of synonyms called synset, in the lexical resource WordNet. It also differentiates between nouns, verbs, adjectives, adverbs. The extension of WordNet is SentiWordNet, which assigns three numerical scores to each synset based on sentiments. They are,

- Positive Measure : PosScore [0,1]
- Negative Measure : NegSocre [0,1]
- ObjSocre : ObjMeasure [0,1]

For better polarity detection we may not dependent only on WordNet lexical relation. Depends on the "part of speech" of the term in a tweet, meanings may have different polarity. Here,we have introduced the Sentiment Classifier that employs Word Sense Disambiguation that uses both WordNet and SentiWordNet lexical dictionaries. Tagging part of speech and tokenization was done on pre-processed tweets. Further, classification is performed applying Word Sense Disambiguation. The goal of WSD approaches is that the true sense of words needs to be identified. The disambiguation process selects the synset (that has maximum similarity score) which finely represents that term in its context, from the WordNet. This is done for each and every word in the selected tweets. The words in tweets are assigned with the sentiment weights by the classifier. Each word is assigned with different weight based on the sense in which the word is used in the tweet. The score on sentiments is calculated for each term in the sentence using the SentiWordNet and the calculation on sentiment score for the sentence is done by summing up the Sentiment sores of each word of the sentence.

4. IMPLEMENTATION AND RESULTS

We have retrieved the tweets about Covid-19 posted during January – June and we have performed Sentiment Analysis on those tweets based on the lexical sources WordNet and SentiWordNet. The Twitter API is used to access the tweets related to a query (Covid-19, Coronavirus in our case) in to a .CSV file. Also we have downloaded the COVID-19 data from TwitterBinder website in to a .CSV file. For using the Twitter API we have got the login credentials such as consumer key, access token. We have extracted around 10000 tweets, and then classified the tweets by calculating the positive / negative sentiment score for these tweets using the lexical sources. The tweets are categorized in to basic sentiment categories positive and negative. Also few samples of tweets classified into positive considering the information on Vaccine, increase in number of samples tested and government policies to address the public concern. Few tweets have been categorized in to negative considering the fear, sadness expressed due to increase in positivity of infection at a fast pace, rise in death rate and decrease in global economy etc. The results are illustrated in the figures 2-5 given below.
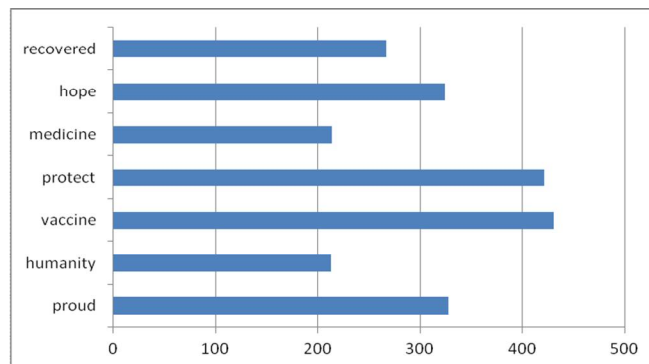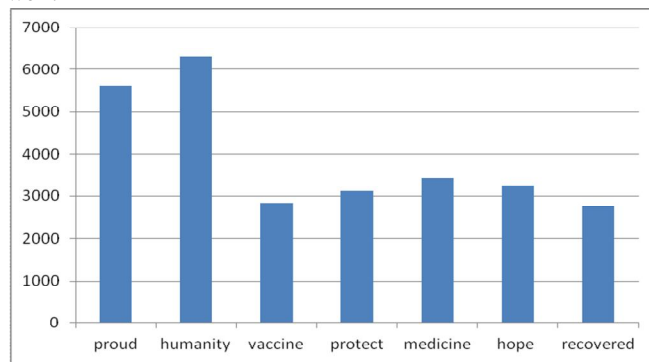

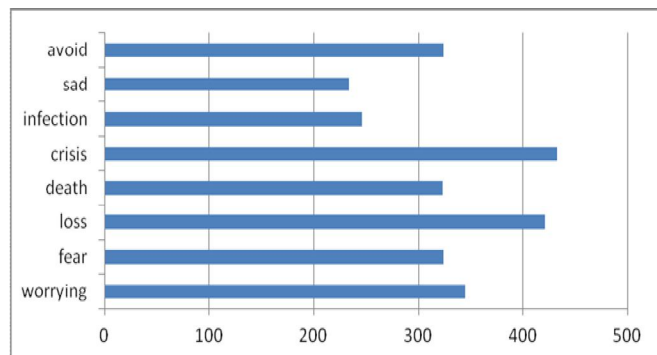**Figure 2:** Classification of Tweets in to Emotion – Trust
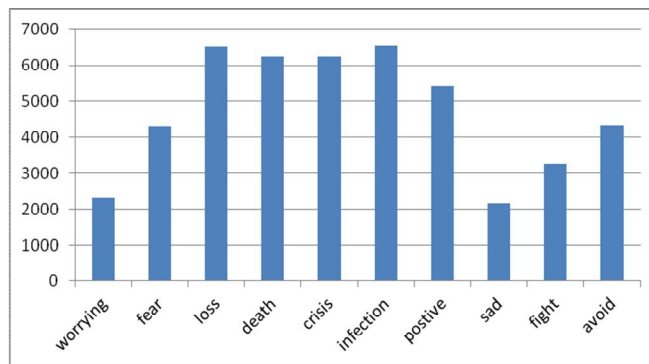
Figure 3: Classification of Tweets in to Sentiment – Negative

The performance of our system is evaluated by considering the accurate classification of emotions that are correctly classified. The accuracy obtained for our work is 83 % which is better than the accuracy obtained in the previous work on emotion analysis during a pandemic. Also we have considered the tweets posted by people across the globe, thus the dataset size considered for classification is very large as well.

Figure 4: Classification of Tweets in to Sentiment – Positive

Figure 5: Classification of Tweets in to Emotion – Fear

5. CONCLUSION

The main goal of this work is to develop a classification model which could categorizes the opinions expressed both implicitly and explicitly in social media platform, Twitter during a pandemic like Covid-19.The tweets extracted in real time using the Twitter API and the raw data got from Twitterminder website were analysed to extract the sentiments to help data analysts to classify into categories of sentiments and emotions as well. Using the experimental results it is clear that when we employ the Word Sense Disambiguation in Sentiment Analysis we could improve the classification accuracy at a reasonable rate than that of Sentiment Analysis without WSD. The findings could be used by the authorities of Government and organization to administer the psychological wellbeing of the people in the society and to initiate the essential measures to address the needs of the people in the community.

REFERENCES

1. Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis", *Foundations and Trends® in Information Retrieval* 2, no. 1–2 (2008): 1-135. https://doi.org/10.1561/1500000011
2. Firmino Alves, André Luiz, Claudio De Souza Bapt ista, Anderson Almeida Firmino, Maxwell Guimarães de Oliveira, and Anselmo Cardoso de Paiva, "A Comparison of SVM versus naive-bayes techniques for sentiment analysis in tweets: a case study with the 2013 FIFA confedations cup", In Proceedings of the 20th Brazilian Symposium on Multimedia and the Web, pp. 123-130. 2014.
3. N. D. Valakunde, Dr. M. S. Patwardhan, "Multi-Aspectand Multi-Class Based Document Sentiment Analysis of Educational Data Catering Accreditation Process", IEEE, International Conference on Cloud and Ubiquitous Computing and Emerging Technologies, 2013.
4. Walaa Medhat a, Ahmed Hassan b, Hoda Korashy b, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal science direct, 2014.
5. Bing Liu, "Sentiment Analysis and Opinion Mining", Morgan and Claypool Publishers, May 2012.
6. A. Giachanou and F. Crestani, "Like it or not: A survey of T witter sentiment analysis methods," *ACM Comput. Surv.*, vol. 49, no. 2, 2016, doi: 10.1145/2938640.
7. H. J. Do, C. G. Lim, Y. J. Kim, and H. J. Choi, "Analyzing emotions in twitter during a crisis: A case study of the 2015 Middle East Respiratory Syndrome outbreak in Kore", 2016 Int. Conf. Big Data Smart Comput. BigComp 2016, pp. 415 − 418, 2016, doi: 10.1109/BIGCOMP.2016.7425960.
8. Dong, R., O'Mahony, M.P., Schaal, M., McCarthy, K., Smyth, B. (2016). Combining similarity and sentiment in opinion mining for product recommendation. *Journal of Intelligent Information Systems*, *46*(2), 285–312.
9. Farhan Hassan Khan, Saba Bashir, Usman Qamar, TOM: Twitter opinion mining framework using hybrid classification scheme Decision Support Systems 57 (2014) 245257, 2013 Elsevier B.V.
10. H. J. Do, C. G. Lim, Y. J. Kim, and H. J. Choi, "Analyzing emotions in twitter during a crisis: A case study of the 2015 Middle East Respiratory Syndrome outbreak in Korea," 2016 Int. Conf. Big Data Smart Comput. BigComp 2016, pp. 415 − 418, 2016, doi: 10.1109/BIGCOMP.2016.7425960.
11. V. Lampos and N. Cristianini, "Tracking the flu pandemic by monitoring the social web," *2010 2nd Int. Work. Cogn. Inf. Process. CIP2010*, pp. 411 − 416, 2010, doi: 10.1109/CIP.2010.5604088.
12. F. T. Giuntini *et al.*, "How do i feel? Identifying emotional expressions on facebook reactions using clustering mechanism," *IEEE Access*, vol. 7, pp. 53909 − 53921, 2019, doi: 10.1109/ACCESS.2019.2913136.
13. A. Bifet, G. Holmes, B. Pfahringer, MOA-TweetReader: real-time analysis in twitter streaming data LNCS 6926, Springer-Verlag, Berlin Heidelberg, 2011, pp. 4660.
14. Lakshmana Phaneendra Maguluri , R Ragupathy, "A New sentiment score based improved Bayesian networks for real-time intraday stock trend classification", International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No.4, July – August 2019, pp.1045-1055. https://doi.org/10.30534/ijatcse/2019/10842019
15. Deepa Mary Mathews , Sajimon Abraham, "Human Annotation and Interpretation of Public Sentiments about Jio Coin marked in Social Networks using Machine Learning Algorithms", International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No.4, July – August 2019, pp.1162-1167. https://doi.org/10.30534/ijatcse/2019/25842019