International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Email:editorijoer@gmail.com http://www.ijoer.in ISSN: 2321-7758

Vol.4., Issue.5., 2016
(Sept.-Oct.)

RESEARCH ARTICLE

ISSN: 2321-7758

ENHANCING THE SENTIMENT CLASSIFICATION ACCURACY OF TWITTER DATA USING PREPROCESSING TECHNIQUES

M.BHUVANESWARI[1], Dr.V.SRIVIDHYA[2]

[1]M.Phil Research Scholar, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore.

[2]Assistant Professor (SS), Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore

ABSTRACT

Nowadays social networking sites are increasing rapidly in this world where text communication plays a major role. There is wide usage of social networking sites among all age groups. Sentiment analysis can be used for business development, reviews about various social activities. Sentiment Analysis is the computational handling of opinions, sentiments and subjectivity of text. Common applications of sentiment analysis include the automatic determination of whether a review posted online (of a movie, a book, or a consumer product) is positive or negative toward the item being reviewed. Thus for sentiment analysis, preprocessing is an essential task. This research paper focuses on the preprocessing techniques to enhance the accuracy of the sentiment classification.

KEYWORDS: Sentiment Analysis, Sentiment Classification, Social media, Twitter, Preprocessing

I. INTRODUCTION

Sentiment analysis or Opinion mining refers to abroad challenging area of NLP, computational linguistics and text mining. It aims to determine the attitude of a speaker or a writer with respect to some topic. Sentiment analysis is a process where the dataset consists of emotions, attitudes or opinions which take into account the way a human thinks [1]. In a sentence, trying to understand the positive and the negative aspect is a very difficult task. The features used to classify the sentences should have a very strong adjective in order to summarize the review.

These contents are even written in different approaches which are not easily deduced by the users or the firms making it difficult to classify them. Sentiment analysis influences users to classify whether the information about the product is satisfactory or not before they acquire it. Marketers and firms use this analysis to understand about their products or services in such a way that it can be offered as per the user's needs.

Sentiment analysis is widely applied to reviews and social media network. Twitter **is** an online social network used to send and read short messages called "tweets".The tweets used to analyze and predict the future directions by public opinion for Polling, stock market, products etc.

Preprocessing is an efficient task in sentiment analysis. Preprocessing is used to remove the unnecessary words in text data. After preprocessing the data set, send as input to the sentiment classification. Then automatically the

accuracy of the sentiment classification will be improved.

II. RELATED WORK

In [2]authors to classify movie reviews using various supervised machine learning algorithms, such as Naïve Bayesian, Maximum Entropy, Stochastic Gradient Descent and Support Vector Machine. These algorithms are implemented using n-gram approach on dataset. It is observed that as the value of 'n' in n-gram increases the classification accuracy. Again, use of TF-IDF and Count Vectorizer techniques are combination for converting the text into matrix of numbers also help to obtain the value of accuracy.

Unstructured text has huge amount information which is not easily used by the computer for processing. So that we require certain techniques to accomplish this task for extracting required patterns. Text mining plays an important role of extracting useful patterns from unstructured text. It is one of the emerging technologies for Knowledge Discovery Process. In this paper, a survey of text mining & its techniques, applications, merits and demerits of text mining have been presented. Text mining technique is basically used for extracting pattern from unstructured data [3].

Web and social network, large amount of data are generated on Internet every day. This web data can be mined and useful knowledge information can be fetched through opinion mining process. This paper discussed different opinion classification and summarization approaches, and their outcomes. This study shows that machine learning approach works well for sentiment analysis of data in particular domain such as movie, product, hotel etc., while lexicon based approach is suitable for short text in micro blogs, tweets, and comments data on web[4].

In [5] Initiating new research questions of analyzing online product reviews and other valuable online information from a domain user's point of view and exploring how such online reviews can really benefit ordinary users. In the case of product reviews there exists a visible gap between the designer's perspective and the domain user's perspective. Also that, not a single classifier can be called completely efficient as the results depend on a number of factors.

Sentiment analysis/opinion mining is play vital role to make decision about product /services. Opinion mining not only encompasses concepts of text mining but also the concepts of information retrieval. Major challenges in opinion mining includes feature weighting which plays a crucial role for good classification[**6**].Authors used an ensemble framework for Sentiment Classification which is obtained by combining various feature sets and classification techniques. In their work, they used two types of feature sets (Part-of-speech information and Word-relations) and three base classifiers (Naive Bayes, Maximum Entropy and Support Vector Machines). They applied ensemble approaches like fixed combination, weighted combination and Meta-classifier combination for sentiment classification and obtained better accuracy [7]

III. METHODOLOGY

Twitter dataset as taken as input to the proposed methodology to preprocess the data. Data preprocessing is done by using various preprocessing techniques to eliminate the unwanted text data in tweets.

The Figure 1 shows the various preprocessing steps involved in this proposed work. The twitter dataset are initially needed to be screened by several steps for sentiment classification.

Pre-processing steps involved in this proposed methodology are

- **Remove Re-tweet Entities:** Retweet is used in the twitter website to show the tweeting content that has been posted by another user. The format is RT @ username where username is the twitter name of the person who is retweeting.
- **Remove @ people name:** It is necessary to remove the @ <people name> in the tweets which offers better results.

- **Remove Punctuations:** Punctuations in tweets are not necessary for sentiment classification. Punctuations removal yield a better results.
- **Remove Numbers:** Numerals are not required for the tweets classification and number removal yields a best performance.

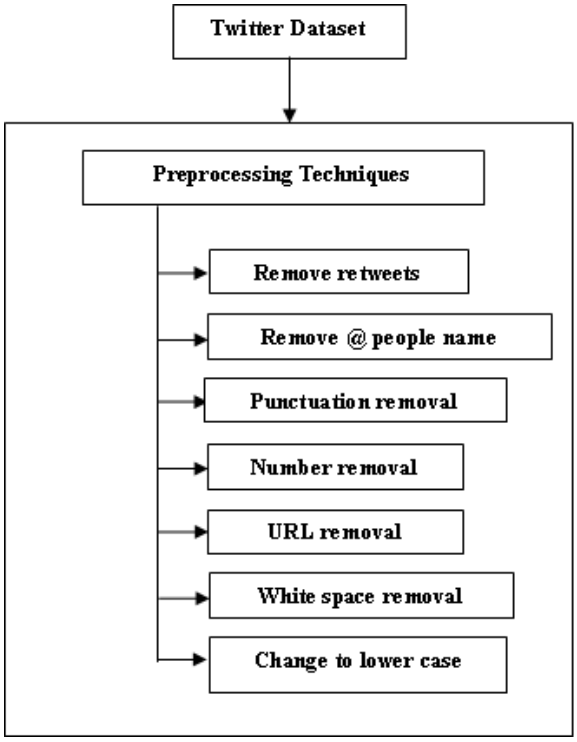

Fig. 1 Preprocessing Methodology

- **Remove URL:** The website links are often attached in the tweet which creates a numerous redundancies which is not necessary for sentiment tweets classification.
- **Remove White spaces:** This step is used to remove the unwanted white space which helps for the tokenization of the tweets.
- **Lowercase:** Converting the tweets to lower case helps the further steps involved in the proposed methodology.

IV. RESULTS AND DISCUSSION

The proposed research work has been developed using R tool. The R tool uses various functions to preprocess the twitter data or tweets. In total 7156 tweets are taken as input for preprocessing the data.

The sample results are obtained from the proposed approach is given in this section. Here 10 tweets are taken as input for sample results. The various preprocessing steps are used for preprocessing the data set, which includes removal of retweet, @people name, punctuation, URL or http link, number removal, white space removal and finally change to lowercase. The removal of retweet result is shown in the following Fig. 2.

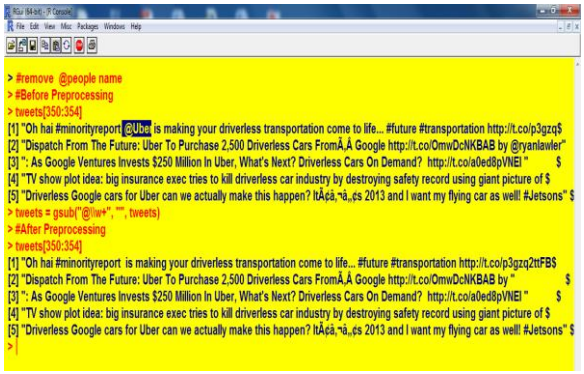

Fig. 2. Removal of Retweets

The following Fig. 3 shows the removal of the @people name or username of the tweets who are posted the tweets in twitter. Before preprocessing and after preprocessing are shown in the above figure.

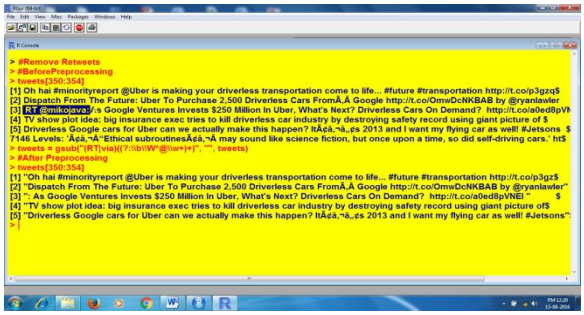

Fig. 3 Removal of @people name

The result of punctuation removal of tweets which is remove the exclamation mark, semicolon, colon, question mark, backslash, dots, etc., are shown in Fig. 4.

The following Fig. 5 shows the result of URL or html link removal of tweets, which is some users are posted a tweet with some html link. Those links are removed. Before preprocessing and after preprocessing are shown in the figure.

The Fig. 6 shows the final result of the preprocessing stage. The input twitter data set preprocessed with several steps and finally it converted to lowercase.

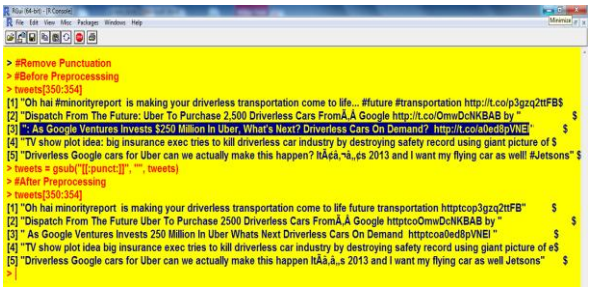

Fig. 4 Punctuation Removal

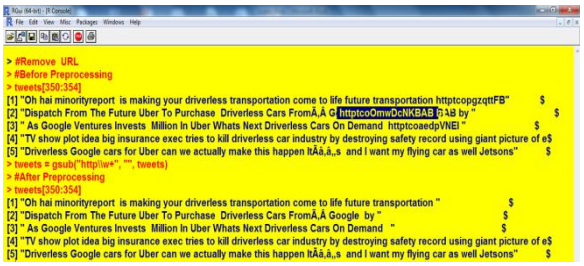

Fig. 5 URL Removal

The final result shown in the following figure.

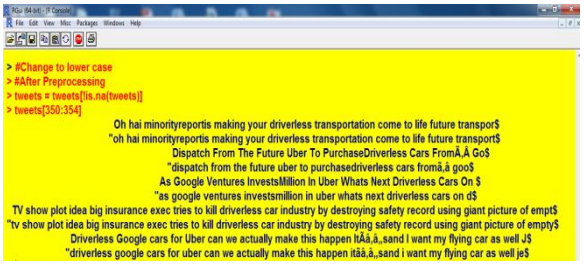

Fig. 6 Final preprocessing result

V. CONCLUSION

In this paper, we discussed various preprocessing techniques for sentiment classification. The preprocessing step has to be done before applying any classification algorithm. This paper consists of various preprocessing tasks, which is retweet removal, @people name removal, URL removal, number removal, punctuation removal, white space removal and finally all tweets are converted to lower case. Now the preprocessed tweets are ready for given as input to any Machine Learning algorithms. The preprocessed tweets are used to improve the sentiment classification accuracy.

REFERENCES

[1]. R. Feldman, " Techniques and Applications for Sentiment Analysis ," Communications of the ACM, Vol. 56 No. 4, pp. 82-89, 2013.

[2]. Abinash Tripathy, Ankitagarwal, santanu kumarrath "classification of sentiment reviews using n-gram machine learning approach", Elsevier publication, Expert system with applications,(2016).

[3]. Amrut M. Jadhav1, Devendra P. Gadekar "A Survey on Text Mining and Its Techniques" International Journal of Science and Research (IJSR) (2012): Volume 3 Issue 11, November 2014.

[4]. Veeramani.S, Karuppusamy.S "A Survey on Sentiment Analysis Technique in Web Opinion Mining" International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Volume 3 Issue 8, August 2014.

[5]. Govindarajan.M ,Romina.M "A Survey of Classification Methods and Applications for Sentiment Analysis", The International Journal Of Engineering And Science (IJES) 2.(12):2013. Page 11-15

[6]. NishaJebaseeli.A .Kirubakaran.E, PhD., "A Survey on Sentiment Analysis of (Product) Reviews", International Journal of Computer Applications (0975 – 888)Volume 47– No.11, June 2012.

[7]. R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," Information Sciences: an International Journal, vol. 181, no. 6, pp. 1138–1152, 2011.

AUTHORS BIOGRAPHY

M.Bhuvaneswari is a Research Scholar, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore. She received Master of Computer Science(M.Sc) degree in 2015 from Bharathiyar University, Coimbatore. Her research interests are Data mining, Computer Networks (wireless Networks), web 2.0 etc.

Dr. V.Srividhya has completed M.Sc., M.Phil., and Ph.D in Computer Science. She is working as Assistant Professor in the Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore. Her fields of research interest are data mining, text mining and Big Data. She has published papers in the international journals and presented research papers in international and national conferences.