Using Slang and Emoticon for Sentiment Analysis of Social Media Data

S. Fouzia Sayeedunnisa[1]
Dept. of IT,
Muffakham Jah Colllege of Engineering
Hyderabad, Telangana State, India

Dr.Nagaratna P Hegde2
Dept. Of CSE
Vasavi College of Engineering
Hyderabad, Telangana State, India

Dr. Khaleel –Ur-Rahman Khan[3]
Dept of CSE
ACE Engineering College
Hyderabad, Telangana State, India

***Abstract* - Social media is a copious source of opinionated data. With the increasing number of people using social media website to vocalize their opinions on various subject, it has become viable to automate these opinions on brand, product, news, story via sentiment analysis aka opinion mining. Opinion mining is gaining insights from these user reviews to know the public view as positive or negative about product, service or brand. This helps business organization and individuals to be proactive in decision making. It finds profound application by helping organization to identify potential product advocates or social media influencers. The current manuscript deals with mining of opinions from the Social media site Twitter on "#Me too" movement. Using tweets generates huge data, which need to be processed and then are to be classified as positive, negative or neutral. The main aim of the manuscript is to use diverse features including emoticons and slang other than conventional Bag of Word features and perform effective Sentiment Classification using these features. We apply a feature selection method to find optimal features and these optimal features are classified. I t is evident from this work that integrating emoticons and slang with conventional Bag of Word model improves the accuracy of classification. The manuscript uses Accuracy, Precision and Recall as the performance metric to analyze the opinion of Twitter users.**

Keywords - Social Network; Twitter, Slang, Emoticons, Bag of Word

I. INTRODUCTION

With the rapid increase in user expressed content over internet has made automatic generation of valuable knowledge from varying documents leading to attraction from public in different segments. Opinion mining or Sentiment analysis refers to a discipline dealing with the analysis as well as the classification of subjective sentiments, opinions, as well as emotions of individuals towards organizations, products, individuals as well as other kinds of topics as pointed out by [3] that are present in text, like tweets as [4] points out, forums [5], reviews [6], news [7], as well as blogs [8]. It is also worth pointing out that sentiment analysis generally makes it possible to identify the trends of individuals as pointed out by [9].

Sentiment Analysis has profound applications; Social media monitoring tools help in performing opinion classification. It can be used by email services to drop spams. One of the trending use of Sentiment Analysis is VOC. It is also used by websites to recommend new content as done by recommender system. Business organizations perform this analysis to determine customer retention on the established and new product [1.].

The organization business pines on its customer satisfaction. Opinion mining helps in customer (VOC) analysis which outcomes profits intensified Sentiment classification faces the challenges of voluminous data available on web which leads the need of feature selection and reduction techniques to have improved performance.

In this paper data is collected from Twitter which is the rising body for bonding people and exchanging valuable information with each other. Twitter helps users to raise concerns or feelings on circumstances of real incidents happening around the world. "#Metoo ", movement was a social phenomenon started in October 2017 about the sexual assault faced by women.

This paper aims in analyzing sentiment of public about the "#Metoo ", movement which flooded the twitter in recent times. Emoticons, slang and Bag of Words are chosen as features for finding the opinion. In Section II the paper further features on Related work, proposed solution in Section III, IV and Conclusion in Section VI.

II. RELATED WORK

Kummer et. Al[5] performed opinion mining using Movie Review dataset with 5000 positive and negative files. The process involved removing features whose overall occurrence frequency is three or less in the corpus. Features that are considered are BOW model with unigram and bigram. For each feature the z score and the confident score is calculated, feature are classified on the confident score obtained from the z score and IG of neighboring bigrams. Achieved accuracy of 85% for unigram and 84% for bigram.

Neethu et. Al[6] analyzed tweets by querying Twitter for electronic products. The data collected comprised 600 positive and negative tweets. They adopted the method of POS tagging the data after preprocessing it and then classified using three Machine L earning classifiers as Naïve Bayes, SVM and Max Entropy. The classification process comprised of an ensemble of these three classifiers using voting rule, which will classify based on the majority output of the classifiers in the ensemble. The ensemble method could achieve an accuracy of 90%.

Asghar et al[7] extracted Sentiments of user review dataset on drug, car and hotel using a rule-based framework.It classifies user reviews by using four classifiers, namely: (i)Emoticon Classifier (EC), (ii) Modifier and Negation Classifier (MNC), (iii) SentiWordNet Classifier (SWNC), and (iv) Domain Specific Classifier (DSC).They could successfully achieve 90% accuracy.

Akhtar et. al [8] executes an ensemble method for classification of sentiment. This model planned feature optimization scheme for lessening the number of features that constructed on PSO. The selection of features is conducted in 3 dimensions so that every aspect reduces the count of features with respect to individual-classifier. Here, the classifiers utilized are Maximum Entropy, SVM & Conditional Random-Field.

III. PROPOSED SOLUTION

Twitter the most popular microblogging site was queried for the phenomenal "#Metoo", movement. This movement gained quite footage on social media involving names of certain big people in the film industry.[] Tweets are short messages of 140 characters encompassing text, slang, photo, emoticons and videos. The messages can be modelled as Bag of Words, unigram, bigram or an n-gram. Each of these cater to features which help in efficient classification of tweets. The first step towards efficient identification of sentiment is preprocessing. It plays a significant role in removal of noise and unwanted text from the tweets. The process further extract features for sentiment classification. The features selected for classification include BOW model, emoticons and slang. The optimal features are extracted by using Mutual Information which is a supervised method representing the correlation between the class and the feature. The low value features are discarded and then classified using SVM. The performance of the classifier is evaluated using Accuracy as the metric.

A. Data Collection

Tweets are extracted from Twitter using the Twitter Search API with the keyword "#Metoo". Comma Separated Values (CSV) file format is used to save the extracted tweets. User generated tweets of only English language was used for analysis. The first step of analysis started with considering the subjective and discarding the objective tweets. The sentiment analyzed comprised of two classes i.e. positive and negative, to train the classifier suitably and to avoid over fitting of data for a specific class same amount of tweets for both classes is used.

TABLE 1
INPUT DATA USED FOR ANALYSIS

Dataset	Positive	Negative
Training Data	750	750
Test Data	250	250

The data collected from Twitter was manually labelled and three fourth data was used for training and one fourth for testing. The CSV data file is processed using Python. Python is used for data analytics and encloses several packages for text data analysis

B. *Preprocessing*

Feature extraction is the major and difficult phase in Social media sites because of usage of emoticons and slang words. To make the analysis better a preprocessing step is mandatory. Preprocessing of tweets improves the classifier performance. The basic step in preprocessing of stop is discarding of stop words. Stop words do not convey any emotion and so are discarded. With the stop words, even the usernames do not convey any sentiment so they are discarded. Further all the tweets are converted to lower case. New lines and punctuation are also removed. The last step of preprocessing is stemming in which a word is converted into a root word. After stemming is done all the redundant words are eliminated from the data.

C. Features

Preprocessing converts the raw dataset into a set of words. The preprocessed data is converted into a Bag of word model which act as one among the feature for classification. In this paper we consider subjective features as well as slang and emoticons which also aid in sentiment classification. In classification methods bag of word is used for text classification where each word with its frequency is considered as a feature

1. Bag of word

Bag of word model is used to implement machine learning algorithms for Sentiment Classification. In this framework a document consist of set tokens or features defines as $\{f_1,\ldots f_m\}$. Let $n_i(d)$ be the number of times f_i occurs in document d. Then each tweet is represented as a vector of features in the BOW model.

2. .Slang words

Slang has become the new trend among the social media users. They are formally defines as small text, words or phrases informally. Usage of slang has become common in writing as well since they take less time. It is an uncontrolled way of expressing the emotions of a person which can be anger or happiness [12].

3. Emoticons

Emoticons have become direct signal for communication among social media users. They are the features to most machine learning algorithms which perform sentiment classification of the data. They comprise of punctuations, letters and numbers. In the current manuscript emoticons

dictionaries [20] are used to replace the emoticon with their description.

D. *Optimal Feature Selection:*

Feature selection plays a vital role in narrowing down feature subset or attributes in predictive analysis. When the training data has high volume it helps best against the curse of dimensionality. It plays an important role in reducing the training time of the dataset. The feature selection method chosen is Mutual Information.

1. *Mutual Information (MI)*

Mutual Information is the measure of mutual dependence between two features which can be words, emoticon or slang. Mutual Information is calculated by considering a term *t* and a class *C*, having probabilities *P(t)* and *P(C) as MI(t,C)*

$$MI(t, C) = \log (p(t,c))/(p(t)\times p(c)) \ldots\ldots\ldots \quad (1)$$

MI calculates the joint probability of term t and Class c together with respect to probability of term t and class c independently. If the term belongs or is associated with class C then the joint probability will be larger else it will be zero if there is no significance relationship between term and class.

IV. CLASSIFICATION TECHNIQUES

The classification is the final step towards determining the opinion of text. The process of classification is defined by the following algorithm.

```
Input: Preprocessed Tweets
Output : Opinion Classified
NL: Negative Lexicon Set
PL: Positive Lexicon Set
EPL: Emoticon Positive Set
ENL :Emoticon Negative Set
SLP: Slang Positive Set
SLN: Slang Negative Set
Function Sentiment Classification( Preprocessed Tweets)
BOW = tokenize(Preprocessed tweets)
 Begin
Score =0
For words in token
      IF word in NL  then
        Score=MI(word, NL)
      Else
           IF word in SLN
           Score= Score + MI(word, NL)
           IF word in ENL
           Score= Score + MI(word, ENL)
          IF word not found in NL,SLN, ENL
          Discard the word.
      Endif
 IF Score> 0
 Optimal Feat= word
Classifier(Optimal Feat, Class)
```

Repeat the same process for PL

The above algorithm computes the optimal features using MI and then pass the optimal features with high MI value to the classifier for training. Every feature is taken and checked if it belongs to lexicon, slang or emoticon. If the word is not found in any one these feature then it is discarded. All the features with low MI Score or MI score less than zero are discarded. All the features are scored according to the MI value computed and then arranged in decreasing order. The top optimal features are used by SVM classifier to classify the tweets as positive or negative. The features considered are diverse features.

TABLE 2
PERFORMANCE OF THE CLASSIFIER WITH ALL FEATURES

Metric	BOW	BOW + Emoticons	BOW+Slang	BOW+ Emoticon+ Slang
Accuracy	0.79	0.81	0.82	0.84
Precision	0.76	0.79	0.79	0.81
Recall	0.72	0.72	0.74	0.79

V. EVALUATION

The data collected for performance evaluation of Sentiment Analysis comprised of 2000 tweets with equal number of tweets for both classes. The tweets were collected for the subject"# Metoo". Only subjective tweets with their emoticons and slang were used for evaluation. The objective tweets are discarded. The performance evaluation with the optimal features using MI is shown in Fig 1. To extract slang we have used SLANGSD[12] dictionary which scores the slang words between -2 to +2, where -2 is strong negative and +2 is strong positive. Considering Slang and Emoticons with improves the performance of classification

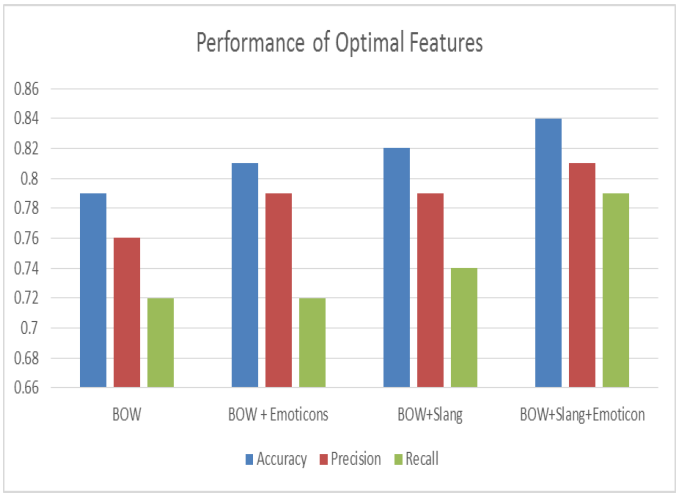

FIG1: PERFORMANCE OF OPTIMAL FEATURES

VI. CONCLUSION

In this paper a framework was developed for sentiment analysis of social media i.e. Twitter was done on two classes i.e. positive and negative. We could achieve 84% Accuracy for binary classification. In this work we used Slang and Emoticon with the rest features and could see an improvement in Accuracy. It is also evident from the results that use of MI as feature selection has reduced the time of processing and showed an improvement in the classification accuracy as well.

In the future we plan to extend this work for finding sentiment of memes and incorporating sarcasm.

VII. REFERENCES.

[1] Tan, W., Blake, M. B., Saleh, I., and Dustdar, S. "Social-network-sourced big data analytics," Internet Computing, IEEE (17:5) 2013, pp 62-69

[2] S. Fouzia Sayeedunissa, Dr. Nagaratna P. Hegde, Dr. Khaleel-Ur-Rahman Khan ,"Using Diverse Feature for Opinion Mining of "Kerala Floods 2018", International Journal of Recent Technology and Engineering Volume -7, Page Number:23-27.

[3] Francisco Villarroel Ordenes[1], Babis Theodoulidis[2], Jamie Burton[2], Thorsten Gruber[3], Mohamed Zaki[4] Analyzing customer experience feedback using text mining: a linguistics-based approach. Journal of Service Research, Volume 17 issue (3),2014, pp. 278-295.

[4] Sayeedunnisa, S.F., Hussain, A.R., Hameed, M.A.: Supervised Opinion Mining of Social Network data using a Bag of word Approach on the cloud. In: Seventh International Conference on Bio-Inspired Computing: Theories and Application (2012)

[5] Bing Liu" Sentiment Analysis and Opinion Mining"ACM Digital library
2012 ISBN:1608458849 9781608458844

[6] Savoy, Olena Kummer—Jacques. "Feature selection in sentiment analysis." Proc. CORIA, no. January (2012): 273-284.

[7] M. S. Neethu and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques," *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, Tiruchengode, 2013, pp. 1-5

[8] Asghar, Muhammad Zubair, et al. "Lexicon-enhanced sentiment analysis framework using rule-based classification scheme." PloS one 12.2 (2017): e0171649

[9] Akhtar, Md Shad, et al. "Feature selection and ensemble construction: A two-step method for aspect-based sentiment analysis." Knowledge-Based Systems 125 (2017): 116-135

[10] Jiménez-Zafra SM, et al. How do we talk about doctors and drugs? Sentiment analysis in forums expressing opinions for medical domain. Artif Intell Med (2018), https://doi.org/10.1016/j.artmed.2018.03.007

[11] Harry Luna-Aveiga1 , J M-Moreira1 , K L-Ortiz1 , O Apolinario , Mario A P-Valverde(&) , M-Zárate , and R V-García "Sentiment Polarity Detection in Social Networks: An Approach for Asthma Disease Management." In: Le NT., van Do T., Nguyen N., Thi H. (eds) Advanced Computational Methods for Knowledge Engineering. ICCSAMA 2017. Advances in Intelligent Systems and Computing, vol 629. Springer, Cham

[12] https://www.kdnuggets.com/2016/09/slangsd-sentiment-dictionaryslang-words.html

[13] https://archive.ics.uci.edu/ml/datasets/OpinRank+Review+Dataset

[14] Liang, Po-Wei, and Bi-Ru Dai. Opinion Mining on Social Media Data. Mobile Data Management (MDM), 2013 IEEE 14th International Conference on. Vol. 2. IEEE, 2013.

[15] Kang, Mangi & Ahn, Jaelim & Lee, Kichun. (2017). Opinion mining using ensemble text hidden Markov models for text classification. Expert Systems with Applications. 94. 10.1016/j.eswa.2017.07.019.

Bing Liu" Sentiment Analysis and Opinion Mining"ACM Digital library
2012 ISBN:1608458849 9781608458844

[16] Saif M. Mohammad, Svetlana Kiritchenko "Using Hashtags to Capture Fine Emotion Categories from Tweets". , Computational Intelligence, Volume 31, Issue 2, Pages 301-326, May 2015.

[17] http://sentiment.nrc.ca/lexicons-for-research/ https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.html

[18] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in Proceedings of the 40th annual meeting on association for computational linguistics, pp. 417–424, Association for Computational Linguistics, 2002..

[19] Celikyilmaz, Asli, DilekHakkani-Tur, and Junlan Feng. Probabilistic model-based sentiment analysis of twitter messages. Spoken Language Technology Workshop (SLT), 2010 IEEE. IEEE, 2010

[20] A. Agarwal ,B. Xie,. I. Vovsha, O Rambow and R. Passonneau "Sentiment Analysis of Twitter Data" ,Proceedings of the Workshop on Language in Social Media (LSM 2011), pages 30–38, Portland, Oregon, 23 June 2011. c 2011 Association for Computational Linguistics

[21] A. Celikyilmaz, D. Hakkani-Tur, and J. Feng, "Probabilistic model-based sentiment analysis of twitter messages," in Spoken Language Technology Workshop (SLT), 2010 IEEE, pp. 79–84, IEEE, 2010.

[22] Y. Wu and F. Ren, "Learning sentimental influence in twitter," in Future Computer Sciences and Application (ICFCSA), 2011 International Conference on, pp. 119–122, IEEE, 2011.

[23] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in Proceedings of LREC, vol. 2010, 2010.

[24] 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.

[25] V. M. K. Peddinti and P. Chintalapoodi, "Domain adaptation in sentiment analysis of twitter," in Analyzing Microtext Workshop, AAAI, 2011.

[26] http://www.netlingo.com

[27] http://techdictionary.com

[28] http://lingo2word.com

[29] http://www.noslang.com

[30] http://www.onlineslangdictionary.com

[31] http://www.smsdictionary.co.uk/

[32] https://stanfordnlp.github.io/CoreNLP