Using Diverse Feature for Opinion Mining of "Kerala Floods 2018"

S. Fouzia Sayeedunnisa, Nagaratna P Hegde, Khaleel Ur Rahman Khan

Abstract*: With the intensive use of social media for communication among people, the outspread of information on these platforms has increased the demand to perform mining of user insights on various topics to gain knowledge .This mining of user reviews to know the positive and negative opinion of people towards a topic, product, brand is Sentiment Analysis. A subject of special interest Sentiment Analysis finds application in various fields viz. Brand Monitoring, Voice of Employee, Social media monitoring .In this paper mining of tweets from Social Media site Twitter is done to analyze the public opinion on "Kerala Floods 2018 ".Twitter generates 500 million messages called 'tweets' per day. Processing these gargantuan tweets is time consuming; this paper aims in processing these tweets into set of features by using Sentiment Lexicons and then applying a filtering method to extract those features which are of high value and discarding all the low value features. Emoticons, slang, hash tags plays a vital role in conversation among people to express their opinions. The extracted features which are huge in number are reduced using the feature selection method Information Gain (IG). The selected high value features through IG are applied to Bayes classifier for classification of opinions. It is apparent from the results that analyzing sentiments using emoticons, slang and hash tag features as one among the features is better than using conventional n- gram features. This manuscript uses Accuracy, Precision, Recall, F- measure and Time for processing to analyze the performance of high value words using IG.

***Index Terms*: Emoticons, n- grams, Slang, Social Network Twitter**

I. INTRODUCTION

Sentiment Analysis is an automated process to classify the opinions of user about a given subject with respect to being positive, negative or neutral[1]. With the use of internet in day to day life and proliferation of social networking sites reflects the countless ways people use these communities and their desires to be a part of this community. In this era, people socialize using micro blogs, which nearly generates 2.5 Qt bytes of data every day. Opinion mining is one prime tool to understand the opinions present in this huge unstructured data. This understanding of opinions helps companies to gain insights and revitalize their brands or product. Sentiment Analysis has various applications; one of the distinguished applications of Sentiment Analysis is building Business Intelligence .Companies can perform Sentiment Analysis on the established and new product to determine the customer retention [2.] A business suspires on its customer's contentment. Opinion mining helps in customer (VOC) analysis which outcomes profits intensified. Sentiment classification faces the challenges of unstructured data which are informal text, consisting of hash tag words, slang, emoticons, elongated words, stop words and opinionated words. This paper highlights the effect of considering slang, emoticons and hash tag words as one among the different features for sentiment classification.

Revised Manuscript Received on December 22, 2018.

S. Fouzia Sayeedunnisa, Department of IT, M.J. College of Engineering and Technology ,Hyderabad, Telangana State, India

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

Dr. Khaleel Ur Rahman Khan, Dept. of CSE, ACE Engineering College, Hyderabad, Telangana State, India

II. RELATED WORK

Saif et al[5] used Senti Circles, a lexicon based method for Twitter, Sentiment Analysis. It analyses the opinion both at entity and tweet level. Every opinionated word has a strength and polarity in a Sentiment Lexicon. Senti Circles updates this strength whenever it finds co-occurring words in different context .At entity level detection this approach shows significant improvement in the Accuracy and F-measure .It achieves a 4-5 % more Accuracy than the SentiStrength approach with Tweet Level analysis. The F-measure of SentiCircles drops by 1% than the baseline approach.

Bhattacharjee et al. [6] performed Sentiment Classification of Telecom Data by mining the World Wide Web. Noise reduction plays a vital role in data extracted from internet, which was achieved by using a lexicon based approach of preprocessing. They used a cosine similarity measure was implemented to categorize opinions between highly negative to highly positive using a five point scale. Naive Bayes, Maximum Entropy and SVM classifiers were used to classify the data which was assigned a cosine similarity value. The cosine similarity classifier was successful in achieving an Accuracy of 82.09% when categorizing the comments in 2 classes i.e. positive and negative. They could engineer the Telecom data in six different categories using the aforesaid technique.

In this Salas-Zárate et.al[7] performed an Aspect Level polarity categorization of Twitter Data on "Diabetes". Their system performed grouping of tweets as positive or negative using a three module approach. In the first module the preprocessing of tweets incorporated stop word removal, normalization, tokenization and POS tagging. In the second step Semantic annotation through Stanford NLP in accordance with the diabetic domain ontology was performed. The third step classified the tweets sentiment by using a proximity approach, and then calculated the polarity of the closest words to the diabetic domain aspect using SentiWordNet (SWN).The performance metric F-measure of 81.24%, Precision of 81.93% and Recall of 81.13% was achieved by the above mentioned technique on data containing 900 tweets about the aspect "Diabetes".

Asghar et.al[8] proposed a "T-SAF: Twitter sentiment analysis framework using a hybrid classification scheme" to enhance the sentiment classification of twitter tweets by integrating four different classifiers. Their classification process comprises of slang classifier for detecting slang, an emoticon classifier for detecting emoticons in the tweets, a SentiWordNet based classifier for detecting the polarity of opinionated words and a Domain Specific classifier to categorize the tweets more precisely. They achieved significant improvement in the performance at sentence and document level using the four tier classification scheme. The authors propose the process of identification sarcasm as a future research direction.

Mallis et. al[9] performed Twitter Opinion Mining of Tweets for different subjects using Hashtags. A Greek Sentiment Lexicon was used for Sentiment of Greek Tweets. This lexicon consisted of the six emotions as :anger, disgust, fear, happiness, sadness and surprise. They could achieve more accurate results for emotions Anger and Happiness.

III. PROPOSED SOLUTION

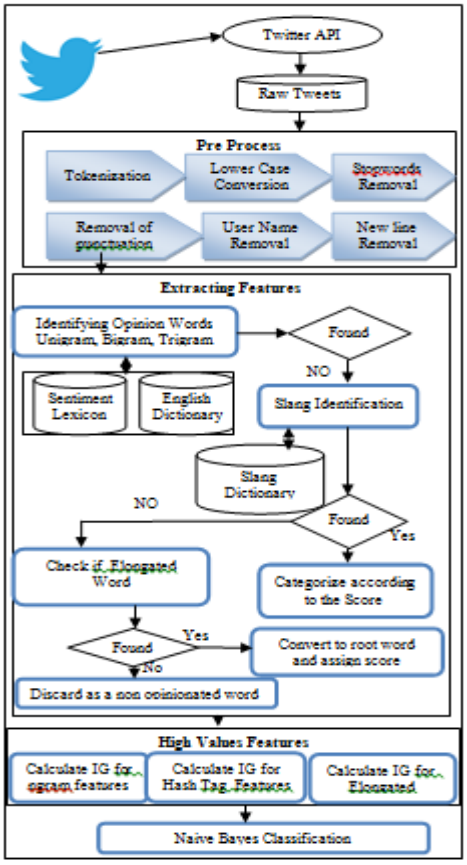

Fig.1 Sentiment Analysis using Distinct Features

The social micro blogging site Twitter was queried using Twitter API to collect tweets regarding "Kerala Floods 2018".This flood was recorded as the most devastating flood experienced in century. A million people were dislocated and over 483 people died [10]. Tweets are messages posted on Twitter comprising text, photos, videos, opinion words, elongated words, slang words, emoticons and also missing words. Each of the aforementioned plays a vital role as distinct features for opinion classification .Identification of Sentiments is done by a four step process . The twitter data contains lot of noise so the first step towards opinion classification is removal of noise and unwanted data through Pre-processing. Features play a key role in identification of opinions, i.e. the second step extracting various features like the ngrams comprising unigram ,bigram and trigram, Hashtag feature, Slang Identification, Emoticon and Elongated features .The next step is selecting high value features using IG and discarding all the low value features . Naïve Bayes Classifier is applied to these concentrated features to evaluate the performance using Accuracy, Precision, Recall, F- Measure and Time taken for processing

A. Data Collection

Twitter Search API is queried with the keyword “Kerala Floods 2018”.The extracted tweets from the Twitter API were collected and saved in a Comma Separated Values (CSV) file format.All foreign language tweets were discarded, extracting the English Language Tweets. Tweets can be further classified as subjective and objective. Objective tweets do not convey any opinion so they are discarded. The emotional measure of tweets is present in subjective tweets, hence used for further processing. To constrain from over-fitting of any specific class, uniform tweets in both classes i.e. positive and negative were chosen to train the model .Our data extracted from twitter consisted of 4000 tweets which was equally divided among both the classes. Python the most used for data analysis is used to implement the system.

B. Preprocessing

Data preprocessing prepares the data for analysis. It assures that the data is clean without noise to be applied to predictive models. The preprocessing includes

- Changing Uppercase to lowercase: To have ease in the feature extraction and selection process, lower case conversion of tweets is done.
- Including Usernames: User names are non-opinionated, so they are discarded in the preprocessing. Text prefixed @ symbol specifies the user name who is directing the message
- New lines removal: New lines just increases the size of the data. They are discarded during preprocessing.
- Removal of punctuations: Punctuations do not convey any sentiment and so as a part of preprocessing they are removed.
- Tokenization: Tweets are short messages ,which has to be tokenized. Branching of these messages into keywords, phrases is tokenization.
- Stop word removal: Text data consist of huge data which do not convey any opinion, i.e. Stop words, one example is connectors (and, the, they ect.).Python has a Natural language tool kit to discard these stop words.

C. Features

A sequence of words, certain emoticons, slang, elongated, Hash tag words are the outcome of preprocessing. Each of this subscribe to a distinct feature. Identification of distinct features dominates the classification of tweets into two classes. They are many features described by Xia et. al[11]. The proposed method incorporates slangs, emoticons, elongated words, hash tagged words and the ngram for opinion classification.

ngram of words

In Text Mining an ***n*-gram** refers to a collection of co-occurring words in any text. The preprocessed tokenized tweets are passed to the feature extraction module. Feature extraction process the ngram tweets into Unigram i.e. size one ngram, Bigram i.e. size 2 n- gram and Trigram i.e. size 3 ngram. The ngrams are compared with the Sentiment 140 Lexicon and NRC Emotion Lexicon [12] to categorize them and label as positive and negative. The following number of ngrams features was collected after preprocessing of 4000 tweets - Unigram 70234, Bigram- 168448 and Trigram -238884 inclusive of both positive and negative sentiment

Hash tag words

Hashtag words are keywords which group vast information pining a theme or topic through these watchwords. The Hashtag used for the current system are "#KeralaFloods". The tweets with these keywords are separated and processed as Unigram Hash and Bigram Hash. NRC Hashtag emotion Lexicon [13] was used to label these Hashtag tweets as positive and negative. The hash tagged tokens collected in 4000 tweets are 40240.

Slang words

Slang is usage of words, phrases, and small text informally. Slangs were more common in speaking rather than writing but with the extensive use of social media it has become equally common in writing. Since they are short forms they are easy to use and take less time. They express the emotional attitude of the person sometimes anger, happiness, disgust in an uncontrolled language.

Elongated words:

One of the very common aspect in use of instant messaging, micro blogging and texting is Elongations also called word lengthening. It is more frequent in these but less in e-mail. In word lengthening a character is repeated two or more times e.g. "crazyyyyy!!!". Elongated words are converted to root words and assigned labels by comparing with the sentiment lexicon. total number of elongated words extracted 7800.

Emoticons

Emoticons have become an integral part among the micro blogging users to express emotional inclination towards a specific topic, brand or event. They consist of letters, numbers and punctuation marks. All the extracted emoticons are replaced by their description. The emoticons which do not convey any opinion are discarded as they just increase the size of data.

D. Feature Extraction

Feature Extraction module extracts the distinct features which are then classified using Naive Bayes. It starts with the preprocessed words which are passed to feature extraction module which identifies the distinct unique feature. The first step in Feature extraction is identifying ngram feature, and label them as positive or negative comparing with the sentiment lexicon [12].If a word is identified as non-opinionated word it is further send to the slang identification module. In this module the word is searched in the SLangSD [14] sentiment dictionary. SlangSD is a sentiment lexicon for slangs which identifies and scores a slang from -2 (strongly negative) to +2 (Strongly positive). All the slang term with zero score are discarded as they are neutral. Slang with score as -2 and -1 are labeled as negative and +2,+1 are labeled as positive. Words not found in the slang dictionary are processed in the elongated word identification module. The Porter Stemmer algorithm is applied to convert the elongated word to root word which is labeled to the respective class after comparing with the standard lexicon set. Word which is not elongated is termed as non-opinionated word and discarded. All the extracted emoticons from the tweets are replaced by their description.

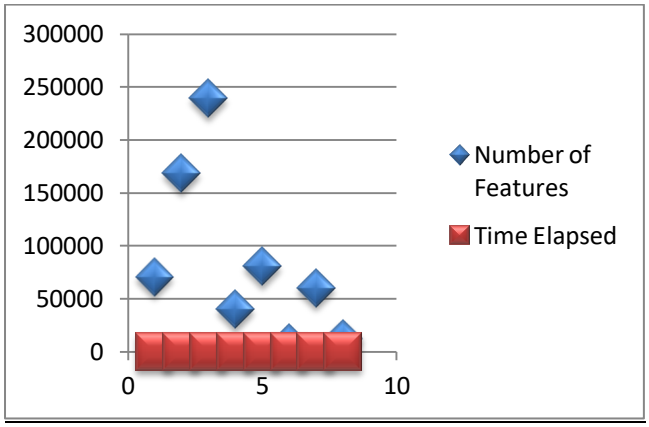

Fig.2 Time Elapsed - Extraction of Diverse Feature Features

E. Feature Subset Selection

Feature Subset Selection *a* part of feature selection is applied to improve the performance of classification. Various feature subset selection measure assigns scores to features. Information Gain (IG), Chi- Square test, LDA, Pearson's Correlation are few feature selection measures. They help in performing better classification by reducing the voluminous features by keeping a threshold value.

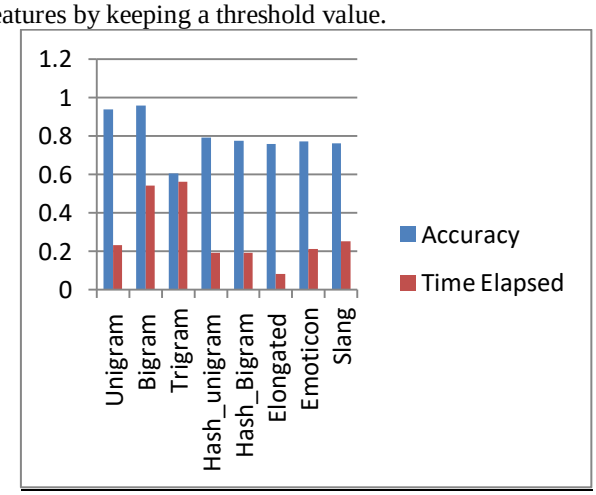

Fig.3 Time Elapsed - Classification of High Value Features

All features with a high IG value are considered to be more informative for sentiment classification. The features with a very low IG value are not considered for classification.

For information gain (IG) we used the Shannon entropy measure [Shannon, 1948] in which:

$$IG(C,A) = H(C) - H(C|A)$$

Where
IG(C, A) information gain for feature A in Class C;
$H(C)=- \sum p(C=i)\log p(C= i)$ entropy across sentiment classes C;
$H(C|A)=-\sum p(C=i|A)\log p(C=i|A)$ specific feature conditional entropy;
H(C) is 1 if the tweets of both classes are equal. The IG range of values for every attribute lies between 0-1, greater values signifying greater IG. Sentiment Features with IG value more than 0.05 are selected

IV. CLASSIFICATION TECHNIQUES

Various Machine Learning algorithms are used for Opinion classification of text. One of the simplest among them for classification is the Naïve Bayes classifier. It requires less training data and is fast .

A. Naive Bayes Classifier

A Supervised Learning method which learns by training with class labels is the Naive Bayes classifier. The classifier considers all the features in the feature vector for classification. It is a binary classifier which works well for the two class problem of classification. The classification is done use the class conditional probability given by

$$P(Y|C_j) = \Pi_{i=1} P(Y_i|C_j) \quad \text{---Eq1}$$

Where 'Y' is the feature vector given by $Y=\{y_1,y_2,....y_m\}$ with Class label C_j .Classification of diverse feature are done using the Naive Bayes classifier. All the high valued features selected using IG are passed to the Bayes Classifier to obtain better Sentiment Classification. The evaluation metric used are Accuracy, Precision, Recall and F-measure. The performance of the classifier before applying IG and after applying IG is noted. Table 1 shows the metrics of preprocessed features. IG is calculated on all the features .The top IG features are selected and passed to Naive Bayes for Classification. The experiment results show a better performing classification process. Performance of High Value feature is tabulated in Table 2

Table 1 Performance of the Classifier with all Features

Metric	UniGram	BiGram	TriGram	Hash UniGram	Hash BiGram	Elongated	Emoticons	Slang
Accuracy	0.80	0.85	0.57	0.75	0.72	0.62	0.73	0.72
Precision	0.82	0.90	0.50	0.80	0.74	0.68	0.85	0.81
Recall	0.80	0.81	0.59	0.73	0.75	0.60	0.66	0.69
F- Measure	0.81	0.86	0.54	0.76	0.74	0.64	0.74	0.74

Table 2 Performance of the Classifier with IG Selected Features

Metric	UniGram	BiGram	TriGram	Hash Unigram	Hash BiGram	Elongated	Emoticons	Slang
Accuracy	0.93	0.95	0.60	0.79	0.77	0.75	0.77	0.76
Precision	0.97	0.97	0.67	0.85	0.80	0.82	0.86	0.77
Recall	0.91	0.93	0.57	0.79	0.77	0.72	0.70	0.77
F-Measure	0.94	0.95	0.62	0.82	0.78	0.77	0.77	0.77

V. EVALUATION

We collected 4000 tweets by querying the Twitter API for the subject "KeralaFloods 2018". The three fourth of the preprocessed data was used for training the classifier and for testing one fourth. To label the data for classification standard lexicon [12] set were used. Only subjective tweets and Hashtag tweets were used for processing. The attainment of evaluation metrics using High Knowledge Features (IG) is outlined in Figure 4. It is conspicuous that considering slang, emoticon and elongated with the ngrams features gives a considerate knowledge of the sentiment. It is also clear that the use of slang and emoticons have reached the commons in the social media

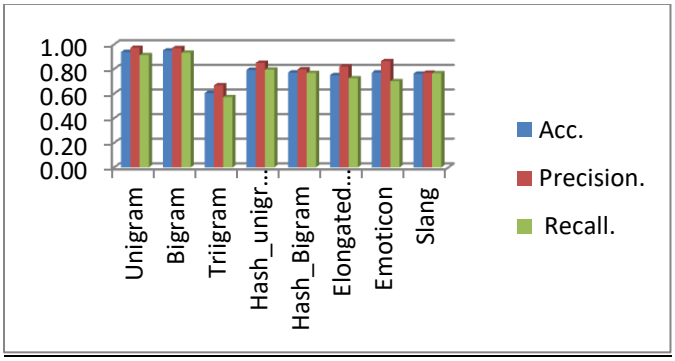

Fig.4 Performance of High Knowledge Features

VI. CONCLUSION & FUTURE WORK

Sentiment Analysis is the future for the data science analysis. It can be done at various levels word, sentence or document level. Recent Literature has shown a significant improvement in Sentiment analysis task at various levels. The process has to have mechanism for efficient handling of sarcasm which can be a future research direction. Our system used diverse feature set including Emoticons, Slang, Elongated words and Hashtag words with the ngrams. It is apparent from the outcomes that slang and emoticon play a significant role in opinion identification. It is also observed that the reduced feature set gives an improved performance metric. Identification of sarcasm and semantic orientation of trigram words can be a research direction for future.

REFERENCES

1. S. Fouzia sayeedunnisa,Dr.Nagaratna P Hegde, Dr. Khaleel Ur Rahman Khan," Sentiment Analysis: Cotemporary Research Affirmation Of Recent Literature"International Journal of Pure and Applied Mathematics Volume 119 No. 15 2018, 1921-1951.
2. 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
3. Manuel, K & Varma Indukuri, Kishore & Krishna, Radha. (2010). Analyzing Internet Slang for Sentiment Mining. 10.1109/VCON.2010.9
4. Francisco Villarroel Ordenes, Babis Theodoulidis, Jamie Burton, Thorsten Gruber, Mohamed Zaki.Analyzing customer experience feedback using text mining: a linguistics-based approach. Journal of Service Research, 17 (3), pp. 278-295
5. Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of Twitter. *Information Processing & Management, 52*(1), 5-19.

6. Saprativa Bhattacharjee, Anirban Das, Ujjwal Bhattacharya, Swapan K. Parui and Sudipta Roy," Sentiment Analysis using Cosine Similarity Measure" IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS) ,2015
7. María del Pilar Salas-Zárate, José Medina-Moreira, Katty Lagos-Ortiz, Harry Luna-Aveiga, Miguel Ángel Rodríguez-García, and Rafael Valencia-García, "Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach," Computational and Mathematical Methods in Medicine, vol. 2017, Article ID 5140631, 9 pages, 2017. https://doi.org/10.1155/2017/5140631.
8. Asghar, Dr. Muhammad & Masud Kundi, Fazal & Ahmad, Shakeel & Khan, Aurangzeb & Khan Saddozai, Furqan. (2017). T-SAF: Twitter sentiment analysis framework using a hybrid classification scheme. Expert Systems. 35. e12233. 10.1111/exsy.12233.
9. Dimitrios Mallis, Georgios Kalamatianos, Dimitrios Nikolaras, Symeon Symeonidis(2015)"Sentiment Analysis of Greek Tweets and Hashtags using a Greek Sentiment Lexicon" The 19th Panhellenic Conference on Informatics (PCI 2015), Athens, Greece, Democritus University of Thrace, DOI:10.13140/RG.2.2.16644.6336
10. https://en.wikipedia.org/wiki/2018_Kerala_floods.
11. 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
12. http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.html
13. http://saifmohammad.com/WebPages/lexicons.html
14. https://www.kdnuggets.com/2016/09/slangsd-sentiment-dictionary-slang-words.html