Abstract

Sentiment analysis is a process of identifying and extracting subjective information in source materials by performing text analysis and Natural Language Processing. It aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. Conventionally, a machine learning algorithm is applied to classify the polarity of a given text into positive, negative or neutral and this classification is done based on emotional states such as ‘angry’, ‘sad’, ‘happy’ etc. A better classification can be achieved by considering emoticons along with emotional states. In many past studies, the emoticons played an important role in building sentiment lexicons and in training machine learning classifiers and also they are considered to be the reliable indicators of sentiment. But, real meaning of all emoticons is not known to many of the social media users. Clustering of words and emoticons in the context of social media will give a good insight about the meaning conveyed by the emoticons. Emoticons are labeled as positive, negative or neutral based on the respective cluster of words they come under. Emoticons are identified from the text and sentiment analysis is performed by using emotional states in the text and emoticons. Such an analysis will result in better classification. This paper focuses on clustering of words and emoticons to know the meaning conveyed by the emoticons and compare the results of sentiment analysis before and after the emoticons are removed from the text.

Key words: Emoticons, Sentiment analysis, clustering.

I. Introduction

Sentiment analysis of a textual data started with comparing the individual words in the text with a bag of positive and negative words and thereby assigning a sentiment score to each word. The score can be +1(positive), -1(negative), 0(neutral). The total score of the textual data gives the sentiment score. This method of sentiment analysis does not help in calculating the exact sentiment in all cases. In order to achieve a better classification of the sentiments the data
analysts started following machine learning that involves supervised classification. A supervised classifier is built based on a training corpus that contains correct label for each input. But many researches on sentiment analysis have focused only on textual information. The social media users tend to express their sentiments in a better way by using emotions and so considering emoticons will lead to more meaningful sentiment analysis. Many past studies have considered emoticons but they have performed sentiment analysis on the basis of real meaning of emoticons. The fact that many social media users does not know the real meaning of emoticons and simply add emoticons to the text by their appearance. Information about how people tend to use emoticons can be gained by clustering the words with emoticons. Thus a better classification can be achieved. This paper aims at comparing the results of sentiment analysis by clustering words and emoticons.

II. Literature Survey

Emoticons are one of the strong indicators of sentiment. But many times the social media users use the emoticons without knowing the real meaning. Clustering of words and emoticons helps in understanding the meaning conveyed by the emoticons [1].

Machine learning plays a major role in sentiment analysis [4]. Before performing machine learning it is required to pre-process tweets. During pre-processing the information in the tweets such as tweet dates, URLs and user names are removed because this information are not required for training the classifier. Also stop words like 'a', 'an', 'the' etc. are removed [4].

In [3], the author talks about the features that are to be considered for sentiment classification. The features considered are:

1. Punctuations
2. n-gram features
3. Pattern based features

In punctuation, features like sentence length in words, number of characters like ‘!’, ‘?’ quotes are used. In word based and n-gram based features unigrams such as ‘good’, ‘bad’, etc. bigrams such as ‘not good’, ‘not bad’ are considered. Sequence like emoticons that contain consecutive punctuation symbols are considered as single word feature. In pattern based features, sentiment hash tags, emoticons are used. The efficiency of the classifier is determined by feature
extractor. In [1] the classification is done using NaiveBayes classifier trained using bag of words model. The classification is done before and after the emoticons is removed from the text. Results have proved that emoticons are better indicators of sentiment. Also in [1] it is also suggested to cluster words and emoticons to better understand the meaning conveyed by those emoticons. The challenge given by this work is to develop an algorithm to perform sentiment analysis based on the clustering results so as to achieve a better classification.

III. Existing System

Conventionally a machine learning approach is used to classify the sentiments of the tweets and during the classification process only textual data is considered. Many past researches that considered emoticons to build the sentiment lexicons, has considered only the real meaning of the emoticons. Sometimes this method of considering real meaning of emoticons does not lead to a robust sentiment classification as people do not know about what the emoticons really mean.

IV. Proposed System

The proposed system mainly performs three tasks

1) Compare the sentence polarity before and after the emoticons are removed from the text.
2) Clustering of words and emoticons to better understand the meaning conveyed by the emoticons.
3) To perform sentiment analysis based on the meaning obtained during clustering.

In order to compare the sentiment polarity with and without emoticons a classifier is trained and tested for two cases:

Before the emoticons removed.
After the emoticons removed.

To cluster the words and emoticons the k means clustering algorithm is used. A classifier is trained to perform sentiment analysis based on this clustering result.

V. System Architecture
Figure 1 shows the architecture of the proposed system. Tweets are retrieved by using a twitter API and it is pre-processed to remove punctuations, white spaces etc. Then it is subjected to the process of classification. During classification the machine is trained and tested to classify the sentiments. During training each tweet in a training corpus that contain correct label for sentiment polarity approximately 21,000 tweets is passed to a feature extractor to generate feature that are required for sentiment classification. The generated feature set is passed to a machine learning algorithm to generate a classifier model. Apart of training corpus is set apart for testing, each tweet in the testing data is subjected to feature extraction to extract features and the generated features are passed to classifier model to predict sentiment polarity labels and the predicted result is tested for accuracy. During prediction, each of the extracted tweet is pre-processed, passed to a feature extractor to generate features and this generated features are passed to the classifier model to produce classified tweets. The process of classification is done before and after emoticons is removed from the tweet to know the importance of emoticons in sentiment analysis. The emoticons play a great role in performing accurate sentiment analysis. However, to consider the meaning of emoticons in the way that users perceive it, the words and emoticons are clustered using k means clustering algorithm and the sentiment classification is done.

VI. Solution Methodology

The steps involved in implementing the proposed system are explained as follows:

A. Extracting tweets

Tweets are extracted with the help of twitter API(Application Program Interface), which is created by signing into twitter developers. Once an API is created a consumer key,a consumer secret,and an access token is generated. These keys are
used for authentication during the extraction of tweets by the secret API. A search API is created to produce tweets by interacting with the twitter API. The tweets produced are saved in csv format.

B. Data pre-processing

Every tweet is pre-processed before sending it into classifier. The tweets are pre-processed to remove urls, to remove white space, replace number of word with word, to remove punctuations, to convert all word into lower case. The pre-processing is done so as to make the classification process easy.

C. Feature Extraction

During this process features that are required for sentiment classification is defined. The success of a classifier is determined by the feature extractor. Some of the features that are considered during feature extraction are:

Words that does not convey any sentiment is filtered out. Repeating letter that is used to stress the emotion is replaced by single letter. E.g. huungrry can be replaced by hungry.

Punctuation such as comma, single or double quotes etc. are removed.

Emoticons such as 😊, 😓, 😐, :-D are considered.

D. Classification

Classification involves three phases:

i) Training

ii) Testing

iii) Prediction.

During training phase, a training corpus is used and training corpus contains correct label for each input tweet. In the context of sentiment analysis, the label defines the sentiment polarity of the tweet. Label can be positive, negative or neutral. Each input tweet in the training data is converter into feature set with the help of a feature extractor. These feature sets capture the basic information about each input that is to be classified. A classification model is generated using a machine learning algorithm whose inputs are pairs of feature sets and labels. During testing phase, each tweet in the testing data is fed to the feature extractor to convert input test tweets to feature sets. The generated feature sets are fed to the model to generate classification labels and the accuracy of the classifier is tested by comparing the classified labels and the predefined labels.
During prediction phase, the extracted tweets are converted to feature sets with the help of some feature extractor. These feature sets are fed to the model and the produced labels are generated.

**E. Testing and Prediction**

This is done before and after the emoticons is removed from the tweet. The Naviebayes classification algorithm is used to perform machine learning.

**F. Sentiment Analysis by Clustering of Words and Emoticons**

The words and emoticons are clustered by using k means clustering algorithm and emoticons that occur with the cluster of positive word are added to the bag of positive words and emoticons that occur with the cluster of negative word are added to the bag of negative words. Then the classification of sentiments is performed as described in the section A.

**VII. Results and Discussion**

![Figure 2. Sentiment analysis with emoticons.](image1)

![Figure 3. Sentiment analysis without emoticons.](image2)
Figure 2 and figure 3 gives clear evidence about the fact that emoticons are the strong indicators of sentiment. Figure 2 shows the results of sentiment analysis with emoticons. Figure 3 shows the results of sentiment analysis without emoticons. It can be inferred that when emoticons are removed from the text, there is a considerable increase in the number of neutral tweets. When emoticons are considered for sentiment analysis, the number of positive tweets, the number of negative tweets increases and the number of neutral tweets decreases. From figure 4 and 5 it can be noticed that, certain tweets are classified as neutral when emoticons are removed. This clearly shows that the emoticons are
strong indicators of sentiment. Also accuracy of the classifier is increased when the emoticons are considered for sentiment analysis.

Further analysis by clustering of tweets and emoticons gave an idea about the meaning of emoticons as users perceive.

The following figure 7 shows some of the words in each of the clusters.

| Cluster1: |
| Emoticons: :) :D :-) |
| Words: great good happy best playful funny happier fantastic lovely wonderful |

| Cluster2: |
| Emoticons: ;) ;-) ;D ;D ;P ;) |
| Words: smile friends music fun naughty kidding pleasure exciting healthy |

| Cluster3: |
| Emoticons: :( /: XD :') (:-( O: ;(-/ :| |
| Words: sad miss broke fuck ugly unexpected unwell problem unhealthy boring don't know worried aching worsened never |

| Cluster4: |
| Emoticons: :P ;D ;P ;] ;p |
| Words: what lol think fun omg great joking bro acting |

| Cluster5: |
| Emoticons: DX |
| Words: music camera smartphone fucking smart |

| Cluster6: |
| Emoticons: :) |
| Words: best fun pleasure friendly funny family playful marvelled smart |

**Figure 7. Word-Emoticon Clusters.**

When sentiment analysis is performed using the word-emoticon clusters, there are some variations in the number of positive and negative tweets. Also there is a slight increase in the accuracy

**VIII. Conclusion and Future Work**

Emoticons play a major role in sentiment analysis. Clustering of words and emoticons could provide a valuable insight into the relationship between emoticons and sentiment polarity. Moreover the words appeared in same clusters comprehended complex meaning conveyed by the emoticons such as :\-\ and ;). An understanding about the importance of emoticons in sentiment analysis could be attained by comparing the sentiments of tweets with and without emoticons. In many cases where the emoticons were the only component in the tweet that expressed some positive or negative sentiment, the sentiment of those tweets became neutral or unclear when the emoticons were removed. Also the accuracy of the classifier became less when the emoticons were removed. A better algorithm to classify the tweets by clustering of words and emoticons can developed so that a better accuracy and more realistic classification can be attained.
An image that is small enough to insert into a text to express an emoticon is called an emoji. Nowadays such kinds of emojis are very popular in social media and each emoji is given a UTF-8 encoding to identify them. But a proper training corpus with emojis is not available to build a classifier model. As a future enhancement emojis can be considered for better sentiment classification.

VIII. References

5. Shuigui Huang, Wenwen Han, Xirong Que and Wendong Wang “Polarity Identification of Sentiment Words based on Emoticons”, 2013.