SENTIMENT ANALYSIS OF MOBILE DATASETS USING NAÏVE BAYES ALGORITHM

Smita Bhanap
Department of Computer Science
Fergusson College, Pune
Pune, India

Dr. Seema Kawthekar
Department of Computer Science & IT
Dr. B.A.M.U. Aurangabad
Aurangabad, India

Abstract: Sentiment Analysis is one of the pursued field of Natural Language Processing (NLP). It is an intellectual process of extracting user's feelings and emotions. The evolution of Internet has driven massive amount of personalized reviews for various related information on the Web specially twitter. These reviews are beneficial for business persons for understanding customer interest, taking better decisions and planning processes. User sentiment refers to the emotions expressed by them through the text reviews. These sentiments can be positive, negative or neutral. The study explores user sentiments and expresses them in terms of user sentiment polarity. Sentiment Analysis poses as a powerful tool for users to extract the needful information, as well as to aggregate the collective sentiments of the reviews. In this paper we present, a lexicon-based approach for sentiment analysis on Twitter. We have used Naïve Bayes algorithm to find sentiment polarities of words in tweeter datasets of some mobile brands. Our approach allows for the detection of sentiment at tweet-level. We evaluate our approach on various mobile brands datasets resulting into accuracy for sentiment polarity classification. We compare various parameters precision, F measure, recall help to improve accuracy.

Keywords: sentiment analysis, Naive Bayes algorithm, accuracy measures.

1. INTRODUCTION

Sentiment analysis is a kind of text classification that catalogs texts based on the sentiment orientation of opinions they contain. It thus plays an important part of Natural Language Processing. NLP is a field of computer science and artificial intelligence that mainly deals with human-computer language interaction. It is the process of detecting the contextual polarity of the text as positive, negative or neutral. The social networks act as a medium where the users post their opinions about an event or a product which are then used for classification. A lot of research work is being held in the field of sentiment analysis due to its significance in the increased competition of current market and also changing needs of the people. Sentiment analysis requires the usage of a training set for its performance, and its quality plays a great role in the accurate evaluation of the text. The semantic analysis of the sentence also increases the meaning and accuracy of the result. [2]. Sentiment Analysis is a sub area of Opinion Mining which aims at inferring the emotion of a user towards a particular item expressed through the reviews. Though sentiments could fall under several classes such as anger, misery, happiness etc, the primary emotions of interest in most cases is the positivity or negativity of the review. Plenty of research has gone into accurate classification of reviews under these categories. [5]

2. LEVELS OF SENTIMENT ANALYSIS

In general, sentiment analysis has been investigated mainly at three levels. In document level the main task is to classify whether a whole opinion document expresses a positive or negative sentiment. In sentence level the main task is to check whether each sentence expressed a positive, negative, or neutral opinion. Document level and the sentence level analyses do not discover what exactly people liked and did not like. Aspect level performs finer-grained analysis. Instead of looking at language constructs (documents, paragraphs, sentences, clauses or phrases), aspect level directly looks at the opinion itself.

3. SENTIMENT ANALYSIS APPROACHES

Sentiment analysis played a great role in the area of researches done by many, there are many methods to carry out sentiment analysis.

- **Machine learning Approach:**

Machine learning strategies work by training an algorithm with a training data set before applying it to the actual data set. Machine learning techniques first trains the algorithm with some particular inputs with known outputs so that later it can work with new unknown data.

- **Lexicon Based Approach:**

Lexicon Based techniques work on an assumption that the collective polarity of a sentence or documents is the sum of polarities of the individual phrases or words. [3]

These approaches can be divided into two methods– Dictionary based method and corpus based method. In dictionary based, it finds the opinion word from review text then finds their synonyms and antonyms from dictionary. The dictionary used may be WordNet or SentiWordNet or other. Corpus-based method helps to find opinion word in a context specific orientation start with a list of opinion word and then find other opinion word in a huge corpus. SentiWordNet 3.0 is most useful dictionary used. It is a lexical resources publically available made up of "synsets" each is associated with a positive, negative numerical score range from 0 to 1. This score is automatically allotted from the WordNet. It uses a semisupervised learning method and an iterative random walk algorithm [4].

4. NAÏVE BAYES ALGORITHM

Naive Bayes classification model works with Bag-of-Word feature extraction which avoids the position of the word in the document. It predicts the probability that a given feature set belongs to a particular label. [1]

5. ACCURACY MEASURES

The accuracy represents the rate at which the method predicts results correctly. The precision is the positive predictive rate, calculates how close the measured values are to each other. The F-measure results refers to the performance of the accuracy. Ideally, a polarity identification method reaches the maximum value of the F-measure, which is 1, meaning that its polarity classification is perfect at it is having highest positive polarity results.

6. EXPERIMENTAL WORK

There is no large public available data set of Twitter tweets with sentiment, so we use Twitter API to collect data. The Twitter API has a parameter that specifies in which language you want to retrieve tweets and we set this parameter to English. We acquire3000 tweets of distinct brands i.e. iphone6, vivo v7, vivo v5, oneplus 5 and Oppo 5. Tweets of these different mobile phones are downloaded using twitter API. Preprocessing of these tweets is done based on removing punctuations, white spaces, converting to lower case. We then used N gram dictionary approach to find polarity. Once the dataset is preprocessed, we partitioned these datasets into train and test datasets. A machine learning algorithm naïve Bayes is applied on test and train datasets. For each dataset, we got accuracy more than 60% which shows that naïve Bayes gives better accuracy.

Table 1: Mobile brands with accuracy measures

Brand	Tweets	Accuracy	Precision	Recall	F measure
Iphone 6	5000	67	0.90	0.75	0.84
Vivo v7	5000	87	1.00	0.62	0.77
Vivo v5	2000	72	0.61	0.63	0.62
Oneplus 5	2000	80	0.95	0.41	0.57
Oppo F5	1000	70	0.61	0.60	0.61

Based on results in Table 1, it is clear that Naïve Bayes gives more accuracy for Iphone 6 Brand as compared to other. In order to improve the accuracy of these datasets using Naïve Bayes we applied various parameters

like precision, F measure and Recall. We got the following results for our mentioned mobile brands. Among all, we got 1.00 precision for Vivo V7 brand which means
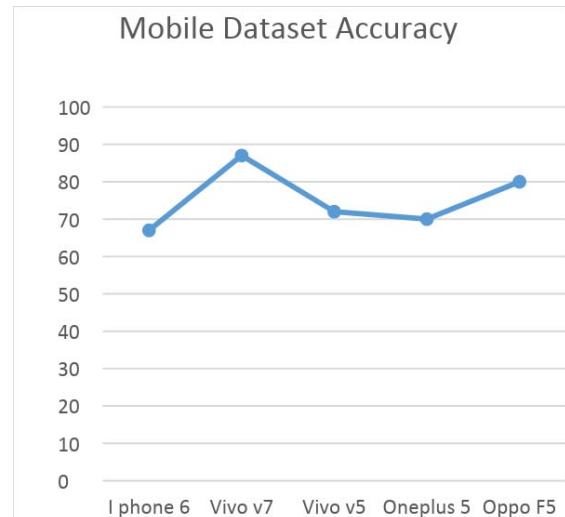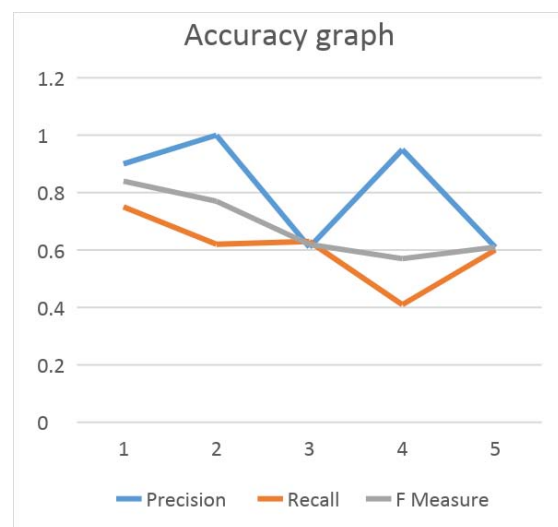

Figure 1: Accuracy chart of mobile datasets

Figure 2: Accuracy graph of mobile datasets

7. CONCLUSION

In this paper we have analyzed mobile brand datasets upto 2000 tweets only. Also we have focused on sentiment polarity of words. Naïve Bayes algorithm is applied to train and test data to find sentiment polarity of overall sentence. After getting accuracy for all datasets we have applied precision, F measure and Recall parameters to it for accuracy improvement. Still there are challenges for handling ambiguous tweets and also work on handling negative polarity levels of tweets. We are going to work on these issues as our upcoming future work. Also comparison of applying various other algorithms is one more issue to be considered to check accuracy.

REFERENCES

1. Amit G. Shirbhate, Sachin N. Deshmukh , "Feature Extraction for Sentiment Classification on Twitter Data" International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064

2. Deepa Anand and Deepan Naorem Semi-supervised Aspect Based Sentiment Analysis for Movies using Review Filtering Procedia Computer Science 84 (2016) 86 – 93

3. M.D. Devika et al. Sentiment Analysis:A Comparative Study On Different ApproachesProcedia Computer Science 87 (2016) 44 – 49

4. Tanvi Hardeniya 1, Dilipkumar A. Borikar Dictionary Based Approach to Sentiment Analysis - A Review International Journal of Advanced Engineering, Management and Science (IJAEMS) [Vol-2, Issue-5, May- 2016] Infogain Publication (Infogainpublication.com) ISSN : 2454-1311

5. Liu B. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers; 2012.