IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 3, Ver. V (May – Jun. 2015), PP 82-86
www.iosrjournals.org

Sentiment Features based Analysis of Online Reviews

Dharmesh Ramani[1], Hazari Prasun[2]
[1]*(Dept of CSE, Parul Institute of Engg. & Tech/ GTU, India)*
[2]*(Dept of CSE, Parul Institute of Engg. & Tech/ GTU, India)*

Abstract : *Sentiment Analysis (SA) and Summarization is a new and emerging field of research which deals with information extraction and knowledge discovery from text using Natural Language Processing and Data Mining technique, which help to track the mood of public about specific products and social or political event. Sentiments of individuals are extremely useful for people and company owner for making several decisions. However decision based upon some of the online review among the large set of review are not very easy for classifying the sentiment. This report introduce new Hybrid Polarity Detection System for SA of short informal text i.e. Twitter post to compare with state-of-art method which used as analysis of Sentiment Summarization. Additionally proposed Hybrid Polarity Detection System derives high performance with new set of features.*

Keywords - *Feature Extraction, Machine Learning Method, Opinion Mining, Sentiment Analysis, Sentiment Classification, Subjectivity Classification, Twitter*

I. Introduction

Today, high measures of informal subjective text content are accessible online with the growing accessibility of social and micro blogging sites. These content or statements are described in several formats, for example news articles, survey and review.

Sentiment Analysis (SA) has recently become the focus of many researchers because of its application and different fields. As it analyzes thought and thought, feelings, attitude, and opinion of individual, analysis of this kind of online review is helpful and demanded for marketing research auditing, public opinion tracking, product reviewing, business research, political review, enhancing of web shopping bases, and so on [6].

Sentiment Analysis is the strategy, used for automatic extracting the polarity of public's subjective opinions from plain natural language statement. Sentiment Analysis is also known as Opinion Mining (OM). Based upon opinion of peoples, anyone can make a decent choice before acquiring any products or items. Sentiment Analysis has an extensive variety of use in e-business, which helps to make good sense of answer of several inquiries like, What do clients think about our items, Which of our users are unsatisfied with the service, What features of our items or product are the worst, Who and how affects our image, What is people in general reaction to some event or some individual [6].

Opinion can be gathered from any person in the world about anything through review sites, surveys, blogs and discussion groups etc [1]. Organizations and product owners who hope to enhance their items/services might emphatically benefit from the rich feedback of clients or users. The most commonly used sources for finding opinion are Blogs, review sites, raw dataset, and Micro-blogging web sites [8].

Online messages that are posted by individual in World Wide Web are generally informal. Analysis or handling of this kind of content is regularly more troublesome if compared with formal writings [4]. The principle difference between formal and informal text is in data preprocessing is formal text often require less preprocessing while informal text often contains emoticons, utilization of bad grammar, sarcasm, and non lexicon- standard words [9]. Therefore, extraction of informal content is regularly more troublesome.

People as often as possible ask their relatives, friends and field masters for recommendation during the decision-making system, and their opinions and point of view are based on experiences and perception. One's perspective around a subject can either be positive or negative, which is term as the polarity detection of the opinion. At the time of sentiment analysis process, it obliges very speedy and concise data so individual can make speedy and exact choice [6]. In sentiment analysis, the information gathered from the reviews has been investigated mainly at three sentiment analysis level [2]:

1.1 Document Sentiment Level

The task at this level is to recognize whether a whole sentiment document expresses a positive or negative sentiment. For example, given an item or product review, the system detects whether the reviews of that item or product communicates an overall positive or negative sentiment about any items. This task is basically term as document-level sentiment classification.

1.2 Sentence Sentiment Level

The task at this level goes to the sentences and figures out if every sentence expressed a positive, negative, or neutral sentiment. Neutral usually characterizes no opinion. This analysis is closely related to subjectivity classification, which perceives sentences as objective sentences, that express real or factual information about the world and subjective sentences that express some individual views, beliefs and emotions. This task of classifying whether a sentence is subjective or objective is terms as subjectivity classification.

1.3 Entity and Aspect Sentiment Level

Above described both the document level and the sentence level do not analyze what exactly individuals liked and did not like. Aspect level serves to derive polarity (positive or negative) and a target of sentiment. A sentiment without its target being recognized is of limited use. Finding out the target of opinion helps to understand the sentiment analysis issue better.

For example, “although the camera quality is not too much great, I still love this mobile”

This statement is positive about the mobile (entity), but negative about its camera quality (aspect). In this way, the goal of this level of examination is to discover sentiments on entities and/or their aspects.

II. Related Work

A lot of research has been carried out via researchers in the sentiment analysis area. Some of the methodologies utilized for sentiment classification are discussed here.

2.1 Naïve Bayes Approach

It is a straight forward and most typically utilized classifier model concentrated around bayes rule that computes post-prior probability of a class concentrated on distribution of words in documents and used for document classification. This methodology work with Bag of Words (BOW) feature extraction which ignore position of words in documents.

The classification approach can be joined with a decision rule, a common rule being, to pick the hypothesis that is most likely which is known as the greatest a posterior model or the MAP decision rule [7].

There are two first order probabilistic models for Naïve Bayes classification are Bernoulli model and the Multinomial model [7]. The Bernoulli model is a Bayesian Network with no word dependencies and binary word features; it likewise produces a Boolean indicator for each one term of the vocabulary depending upon its presence or absence; thus how, the Bernoulli model also considers words that do not appear in the document into record [7]. The Multinomial model is a unigram language model with integer word counts and when the frequency of a word occurring in a document counts; so, a binarized version Of the Multinomial model is utilized which only takes in to account the presence of a word but not its frequency [7]. It is analyze that the multivariate Bernoulli performs well with small vocabulary sizes, however the multinomial model basically performs even better at larger vocabulary sizes, providing on an average 27% decrease in error over the multivariate Bernoulli model at any vocabulary size [7].

2.2 Maximum Entropy

Maximum entropy classification (MaxEnt, or ME) is a feature-based [5] probability distribution estimation model and an alternative technique which has proven effective in a number of natural language processing applications.

Principle of maximum entropy is if not much is known about the data or information, distribution should be as uniform as possible [7]. Significantly, unlike Naive Bayes, MaxEnt makes no assumptions about the relationships between features, and so might potentially performs better when conditional independence assumptions are not met [3]. This implies it should allow adding features like bigrams and phrases to MaxEnt without worrying about its feature overlapping [5]

2.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) is another popular high margin statistical classification technique proposed for sentiment analysis and highly effective for text categorization [3].

The main principal underlying SVM for sentiment classification is to discover a hyper plane which separates the documents as per the sentiment, and the margin between the classes being as high as possible; it also focused around the Structural Risk Minimization principle [7].

Feature selection is an important task in machine learning methods; there are numerous features that must be considered for text classification, to stay away from over fitting and to increase general accuracy [7]. SVM have the potential to handle large feature spaces with high number of measurements.

To deal with a large number of features, traditional text categorization methods assume that some of the features are insignificant, however even the lowest ranked features according to feature selection methods contain considerable information; considering these features as irrelevant often result in a loss of data [7]. Thus how the information loss can be minimized as SVMs does not requires at the time of making an assumption.

Though SVM outperforms all the traditional techniques for sentiment classification, it is a black box technique [7]. It is difficult to research the model of classification and to distinguish which words are more important for classification. This is one of the disadvantages of utilizing SVM as a technique for document classification [7].

III. A Hybrid Polarity Detection System

Modules contain in the Existing Hybrid Polarity Detection System are demonstrated as follow [9]:
3.1 The Preprocessing Module
3.2 Sentiment Feature Generator Module
3.3 Machine Learning Classifier

3.1 Pre-Processing of Data set :

Several Pre-Processing Steps for the Sentiment Summarization of the given data set is taken, this several Steps are:

- @username – removed the username and because these are not considers for sentiments.
- URLs – delete all string that describes links or hyperlinks.
- #hashtag **-** hash tags can give some helpful information, so it is helpful to replace them with the actually same word without the hash. E.g. #Dissertation replaced with Dissertation.
- The target (of sentiment) word is replaced by “TARGET”
- Lower Case – changed over all the content in a string to lower case.
- Stop words **-** a, an, is, the, with and so on, that don't demonstrate any sentiment and can be removed.
- Punctuations and additional white spaces – removed punctuation at the begin and closure of the tweets. E.g.: ‘today is my presentation.!‘ Replaced with 'today is my presentation'.
- Words must begin with an alphabet – deleted each one of those words which don't begin with an alphabet, for example 24th, 7:45pm

3.2 Sentiment Feature Generator Module [9]

Several Features include in the Hybrid Polarity Detection System are as shown in the table, Measurements of all this features are required for further calculation.

Table 1: Features Utilized as a Part of Existing System

F1	Document (or tweet) overall sentiment score using the unsupervised polarity detection algorithm
F2	Number of positive words
F3	Number of negative words
F4	Number of negation words
F5	Number of negation words followed by a positive word
F6	Number of negation words followed by a negative word
F7	Inverse sentiment
F8	Number of positive words followed by target
F9	Number of negative words followed by target
F10	Number of negation words followed by target
F11	Number of positive words followed by a negative word
F12	Number of negative words followed by a positive word
F13	Number of target words followed by a positive word
F14	Number of target words followed by a negative word

3.3 Machine Learning Classifier [9]

Sentiment Summarization of a linear SVM that takes as input the feature set described in the previous subsection that contain opinion about some entity of interest and accordingly classifies tweets (documents) and generate summary of all input tweets.

Now, the proposed approach from the above three module is done by adding two more feature with doing sentiment analysis on live twitter data set. This proposed approach is as shown here:

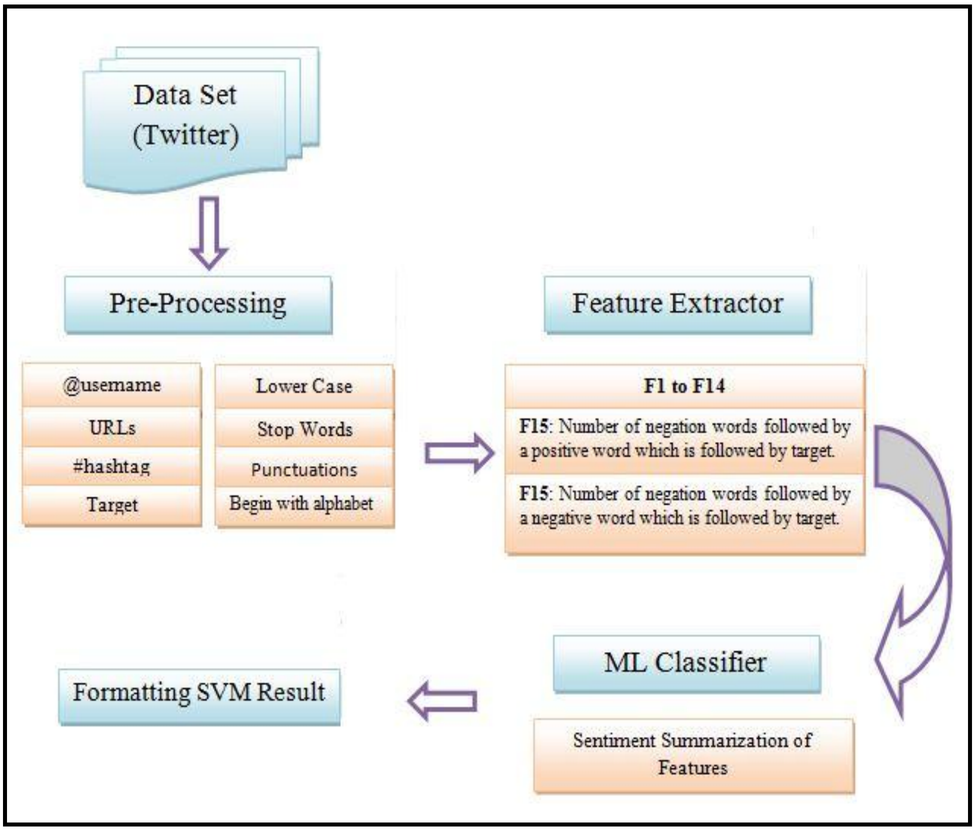
Data Set
(Twitter)
Pre-Processing
@username
URLs
#hashtag
Target
Lower Case
Stop Words
Punctuations
Begin with alphabet
Feature Extractor
F1 to F14
F15: Number of negation words followed by a positive word which is followed by target.
F15: Number of negation words followed by a negative word which is followed by target.
ML Classifier
Sentiment Summarization of Features
Formatting SVM Result

Fig. 1 Proposed Hybrid Polarity Detection System

- **Formatting SVM Result**

Machine Learning Classifier generates set of features with indicating number of count from which the SVM formation done to derive the accuracy of the features set of proposed work.

- **Benefits:**

- All that features proposed in Hybrid System require a very short time to be computed.
- Additional set of Features will help to improve accuracy.

IV. Experimental Result

Here dataset consist of 180-220 Online tweets of different domain like Movie, Hotel and Mobile Product. Further it divided into several Movies, Hotels and Mobile Products.

Now, to evaluate the single class and overall accuracy, we perform

$$\text{Single Class Accuracy} = TP / (TP+FP)$$
$$\text{Overall Accuracy} = (TP+TN) / (TP+FP+TN+FN)$$

Where TP, FP, TN, FN are the number of True Positive, False Positive, True Negatives, False Negatives.

Table 2: the Performance of Proposed hybrid Approach with Compare to Existing Approach

Dataset	Positive Class	Negative Class	Accuracy of Existing Hybrid Approach	Accuracy of Proposed Hybrid Approach
Mobile	65	63	66.98	**67.64**
Movie	34	89	62.11	**62.73**
Hotel	68	86	77.98	**78.54**

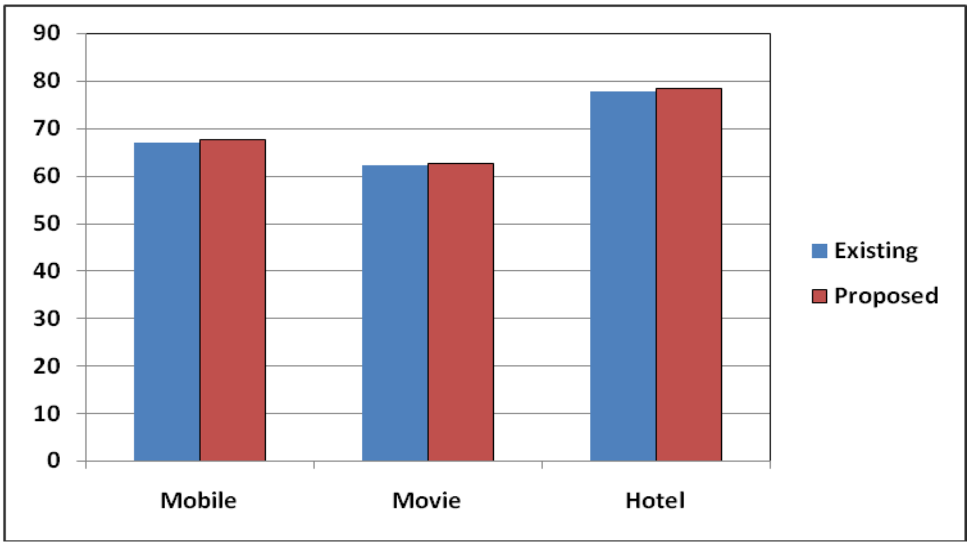

Fig.2: Comparison of Existing Approach and Proposed hybrid Approach

V. Conclusion

As Sentiments of individuals are extremely useful for people and company owner for making several decisions, introduced proposed Hybrid Polarity Detection System for Sentiment Analysis and summarization that uses new set of features, tries to improve the accuracy compare to state-of-the-art techniques to get the clear idea about the marketing research auditing, public opinion tracking, product reviewing, business research, political review, enhancing of web shopping bases, and so on. As per our experiment, we believe that as the part of Sentiment Analysis, Moving towards Sentiment Features rather than manual text processing would be a promising outcome to these issues.

Now, finding more features set that could help to improve the accuracy and also detection of sarcasm would be future work of this study.

References

[1]. Khairullah Khan, Baharum B. Baharudin, Aurangzeb Khan and Fazal-e-Malik, Mining Opinion from Text Documents: A Survey, 3rd IEEE International Conferences on Digital Ecosystem and Technology, 2009.

[2]. Bing Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.

[3]. Bo Pang, Lillian Lee and Shivakumar Vaithyanathan, Thumbs up Sentiment Classification using Machine Learning technique, proceedings of EMNLP,2002, 79-86.

[4]. Khairullah Khan, Baharum Baharudin, Aurnagzeb Khan and Ashraf Ullah, Mining opinion components from unstructured reviews: A review, Journal of King Saud University – Computer and Information Sciences, 2014.

[5]. Alec Go, Richa Bhayani and Lei Huang, Twitter Sentiment Classification using Distant Supervision, CS224N project report, Stanford, 2009, 1-12.

[6]. Dharmesh Ramani and Prasun Hazari, A survey: Sentiment Analysis of Online Reviews, IJARCSSE, 4(11), Nov 2014.

[7]. Sagar Bhuta, Uehit Doshi, AvitDoshi and Meera Narvekar, A Review of Techniques for Sentiment Analysis of Twitter Data, Issues and Challenges in Intelligent Computing Technique (ICICT), 2014, 583-591.

[8]. Blessy Selvam and S.Abirami, A survey on Opinion Mining Framework, International Journal of Advanced Research in Computer and Communication Engineering, 2013.

[9]. Seyed-Ali Bahrainian and Andreas Dengel, Sentiment Analysis using Sentiment Features, IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT), 2013.