ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE) Vol. 2, Special Issue 11, March 2016

Automatic Text Sentiment Analysis Using Naïve Bayes Algorithm

S.Aishwarya[1], P.Vaishnavichandravadhana[2], Mr.S.Chidambaram[3]
[1,2,3] National Engineering College
aishwaryaitnec@gmail.com[1], emailvadhana@gmail.com[2], chidambaramraj1@gmail.com[3]

Abstract- Sentiment Analysis or opinion mining is the computational study of people's opinions, appraisals and emotions towards entities, events and their attributes. In many cases, opinions are hidden in long forum posts and blogs. It is difficult for a human reader to find relevant sites, extract related sentences with opinions, read them, summarize them, and organize them into usable forms. Automated opinion discovery and summarization systems are thus needed. In the past few years, it attracted a great deal of attentions from both academia and industry due to many challenging research problems and a wide range of applications. Language aggression is one among the crimes which could not be supervised clearly so far. In this proposed system the main concern is in detecting the language aggression.

Keywords— Sentiment analysis;Natural language processing; Machine learning; Website reviews; Naive Bayes classification.

I.INTRODUCTION

The research in Text mining field started with sentiment and subjectivity classification, which treated the problem as a text classification problem. Sentiment classification classifies whether an opinionated document (e.g., product reviews) or sentence expresses a positive or negative opinion. Subjectivity classification determines whether a sentence is subjective or objective. Many real-life applications, however, require more detailed analysis because the user often wants to know what the opinions have been expressed on. For example, comment on a person (say A), one wants to know whether A has been praised or criticized.

Let us use the following comment on person A as an example to introduce the general problem (a number is associated with each sentence for easy reference): "(1) I bought a book yesterday. (2) It was such a nice book as you had A. (3) your knowledge about books was awesome. (4) Your advice is adorable. (5) But, your taste on music is really awkward. (6) How could you have such a mean taste on music? "

The first thing that is been noticed is that there are several opinions in this comment. Sentences (2), (3) and (4) express three positive opinions, while sentences (5) and (6) express negative opinions. Then we also notice that the opinions all have some targets on which they are expressed. The opinion in sentence (2) is on book that A has as a whole, and the opinions in sentences (3) and (4) are on the "knowledge" and "advice" of person A respectively. The negative opinion in sentence (5) and (6) is on the taste of person A. This is an important point. Finally, we may also notice the sources or holders of opinions. The source or holder of the opinions in sentences (2), (3) and (4) is the friend of the person A. With the following example, the process flow can be understood. The project gets started with the opinion target.

A. Object and feature:

In general, opinions can be expressed on any target entity, e.g., a product, a service, an individual, an organization, or an event. The term object is used to denote the target entity that has been commented on. An object can have a set of components (or parts) and a set of attributes (or properties), which can be collectively called as the features of the object.

A particular brand of cellular phone is an object. It has a set of components (e.g., battery and screen), and also a set of attributes (e.g., voice quality and size), which are all called features. An opinion can be expressed on any feature of the object and also on the object itself. For example, in "I like that phone. It has a great touch screen", the first sentence expresses a positive opinion on "phone" itself, and the second sentence expresses a positive opinion on its "touch screen" feature.

B. Snippet:

A snippet is a small text segment around a specified keyword in a given document. The text segment can be defined by sentence boundaries, or the number of words. In general, snippets are built around core keywords. Snippetization is important for analyzing web content, since web contents often are noisy.

C. Opinion holder:

The holder of an opinion is the person or organization that expresses the opinion. In the case of product reviews and blogs, opinion holders are usually the authors of the posts. Opinion holders are more important in news articles because they often explicitly state the person or organization that holds a particular opinion.

D. Opinion and orientation:

An opinion on a feature f (or object o) is a positive or negative view or appraisal on f (or o) from an opinion holder. Positive and negative are called opinion orientations.

With these concepts in mind, we can define a model of an object, a model of an opinionated text, and the mining objective, which are collectively called the feature-based sentiment analysis model.

E. Model of an object:

An object o is represented with a finite set of features, $F = \{f_1, f_2 \dots f_n\}$, which includes the object itself as a special feature. Each feature $f_i \in F$ can be expressed with any one of a finite set of words or phrases $W_i = \{w_{i1}, w_{i2}, \dots, w_{im}\}$, which are synonyms of the feature.

F. Model of an opinionated document:

A opinionated document d contains opinions on a set of objects $\{o_1, o_2 \dots o_r\}$ from a set of opinion holders $\{h_1, h_2, \dots, h_p\}$. The opinions on each object o_j are expressed on a subset F_j of features of o_j. An opinion can be either one of the following two types:

1. **Direct opinion**: A direct opinion is a quintuple $(o_j, f_{jk}, oo_{ijkl}, h_i, t_l)$, where o_j is an object, f_{jk} is a feature of the object o_j, oo_{ijkl} is the orientation of the opinion on feature f_{jk} of object o_j, h_i is the opinion holder and t_l is the time when the opinion is expressed by h_i. The opinion orientation oo_{ijkl} can be positive, negative or neutral.

2. **Comparative opinion**: A comparative opinion expresses a preference relation of two or more objects based on some of their shared features. It is usually conveyed using the comparative or superlative form of an adjective or adverb, e.g., "Coke tastes better than Pepsi".

G. Objective of sentiment analysis on direct opinions:

Given an opinionated document d,
1. Discover all opinion quintuples $(o_j, f_{jk}, oo_{ijkl}, h_i, t_l)$ in d, and
2. Identify all synonyms (W_{jk}) of each feature f_{jk} in d.

II. LITERATURE REVIEWS

Bo Pang and et al. [1] ,created a project which lead to statistically significant improvement in polarity classification accuracy. A Kowcika and others [2], Collected useful information from the twitter website and efficiently perform sentiment analysis of tweets regarding the Smart phone war.

Keke Cai and et al. [3], detected the most significant topics hidden behind each sentiment category using a combined PMI and word support metrics. Kiran Shriniwas Doddi and et al. [4] ,provided a platform that filters out negative articles. Richard Colbaughand and et al. [5],presented a project which requires no labeled training documents and is able to provides accurate text classification using only a small lexicon of words of known sentiment/emotion. Apoorv Agarwal and et al. [6], reported an overallgain of over 4% for two classification tasks: a binary,positive versus negative and a 3-way positive versus negative versus neutral. Efthymios Kouloumpis et al. [7] , investigated the utility of linguistic features for detecting the sentiment of Twitter messages. Bing Liu [8] presented a project which detects the opinion spam. Namrata Godbole et al. [9] ,did a project on Sentiment analysis of news and blogs. Amit Gupte et al. [10], presented a project in which efficient algorithms were sorted out.

III.NAIVE BAYES ALGORITHM

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods. Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

- P(c|x) is the posterior probability of class (target) given predictor (attribute).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

IV. PROPOSED METHODOLOGY

In the proposed method, the text corpus acts as an input and the sentiment identification is to be done then the sentiment orientation classification will be done. The text is analyzed whether it has some opinion to be expressed. The expressed opinion will be decided as the sentiment of the corpus. The aggressive words are detected if in case were uttered. Otherwise the positivity of the corpus will be finalized as the summarized output. The above process will be done using the Naïve Bayes classifier algorithm. The following are the steps to be followed to perform the automatic text sentiment analysis, referring to Figure 1

Step 1, **Input**: The website is specifically designed in which the users comment on specific posts. The input is collected from users' reviews of the designed website.

Step 2, **Parser Module**: A natural language parser works out the grammatical structure of sentences, for instance, words which are grouped together are "phrases". During parsing the delimiters are ignored and only the keywords are extracted.

Step 3, **Creation of Training Dataset**: Creating dictionary of words along with their sentiment values. The word dictionary is updated by adding new words which is different from that of the training data set.

Step 4, **Apply Naïve Bayes**: In this the machine learning algorithm called Naïve Bayes algorithm is applied to the input testing dataset along with the training data set. The following is the proposed architecture of automatic text sentiment analysis.

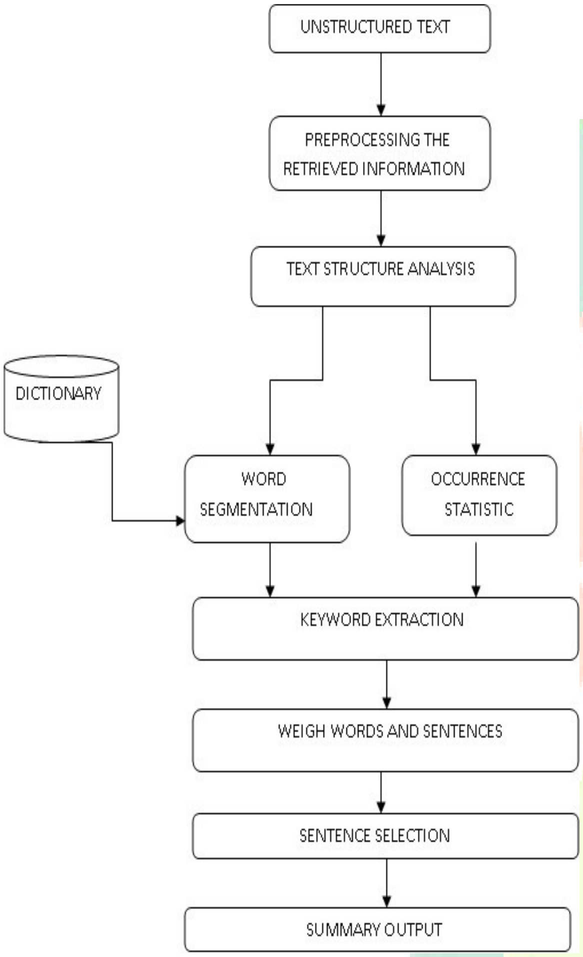

Fig.1. Automatic Text Sentiment Analysis

V. EXPERIMENTAL RESULTS

According to the Bayesian classification model, the testing dataset is analyzed along with the training dataset and the results are obtained. Each and every post in the designed website is analyzed separately and the results are displayed in tabular format representing the positivity and negativity of comments against the post.

The main parameter measure chosen is the accuracy. The Fig. 2 represents the accuracy of the final results. Each post has its own reviews, so the experimental results and the manual results are represented in the table.

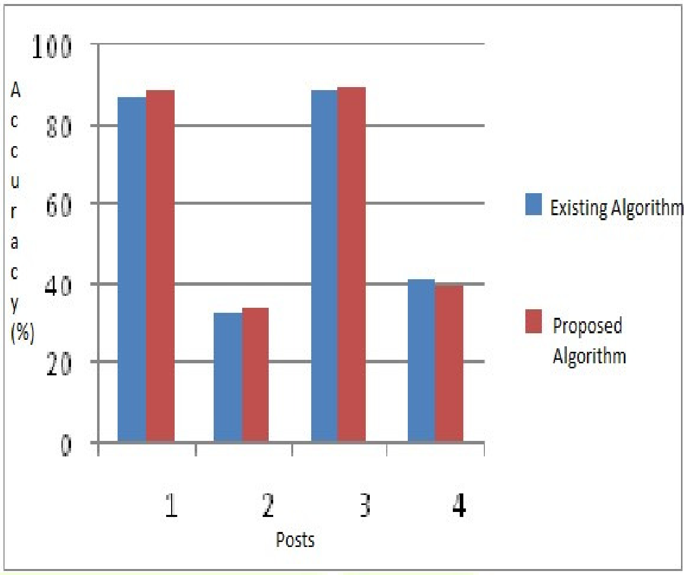

Fig.2. Performance Analysis

VI. CONCLUSION AND FUTURE WORK

In this paper we have implemented Naïve bayes algorithm and successfully classified the reviews of the users against the post positive and negative. The naïve bayes algorithm will be applied to the testing data set along with the training data set. The final output would be in a tabular column describing the amount of positivity and negativity of reviews will be displayed. The accuracy obtained is about 88%. The future work of the system would be inclusion of feature based classification.

Total no. of words	True Positive	False Positive	Accuracy (%)	Error rate (%)
235	135	100	57.46	42.54
536	365	171	68.09	31.91
324	254	100	71.71	28.29
445	425	20	95.5	4.5
521	516	5	99.04	0.96
640	635	5	99.21	0.79

Table 1. Distribution of Positive and Negative Reviews

The feature based classification of the reviews will be considered as the future work of our project.

REFERENCE

[1] Bo Pang ,Lillian Bee , A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts(IIEE), 2014.

[2] Asst. Prof. A Kowcika, Aditi Gupta , Kartik Sondhi, Nishit Shivhre, Raunaq Kumar , Sentiment Analysis for Social Media(IJARCSSE), 2013.

[3] Keke Cai, Scott Spangler!, Ying Chen, LiZhang, Leveraging Sentiment Analysis for Topic Detection(IEEE), 2008.

[4] Kiran Shriniwas Doddi1, Dr. Mrs. Y. V. Haribhakta2, Dr. Parag Kulkarni,Sentiment Classification of News Articles (IJCSIT), 2014.

[5] Richard Colbaugh, Kristin Glass (2013),„Analyzing Social Media Content for Security Informatics'(IEEE) © 28-31,IEEE. pp. 8341-8346, September 2006 .

[6] Apoorv Agarwal Boyi Xie Ilia Vovsha Owen Rambow Rebecca Passonneau, ,Sentiment Analysis of Twitter Data(IJCSIT), 2014.

[7] Efthymios Kouloumpis,Theresa Wilson, Johanna Moore , Twitter Sentiment Analysis:The Good the Bad and the OMG!(AAAI), 2010.

[8] Bing Liu , Sentiment Analysis: A Multi-Faceted Problem(IEEE), © 18-21 August 2005 IEEE. pp. 2341-2346, 2010.

[9] Namrata Godbole, Manjunath Srinivasaiah, Steven Skiena,Large Scale Sentiment Analysis for News and Blogs(IJCSIT),2014

[10] Amit Gupte, Sourabh Joshi, Pratik Gadgul, Akshay Kadam ,Comparative Study of Classification Algorithms used in Sentiment Analysis (IJCSIT),2014.

[11] Beatrice Santorin "Part-Of-Speech Tagging Guidelines For The Penn Treebank Project (3rd Revision)", University Of Pennsylvania.

[12] Dhaval Thakker, PA Photos, UK, Taha Osman,"GATE JAPE Grammar Tutorial", Nottingham Trent University, UK, Phil Lakin, UK, Version 1.0