International Journal of Emerging Trend in Engineering and Basic Sciences (IJEEBS)
ISSN (Online) 2349-6967 *Volume 2 , Issue 1(Jan-Feb 2015), PP1-5*

Sentiment analysis on news articles using Natural Language Processing and Machine Learning Approach.

Pranali Chilekar[1], Swati Ubale[2], Pragati Sonkambale[3], Reema Panarkar[4], Gopal Upadhye[5]

[1 2 3 4 5]*(JSPM's Rajarshi shahu college of of engineering/ Savitribai Phule Pune University ,India)*

Abstract :- Sentiment analysis basically aims at determining the attitude of a writer with respect to some topic or the overall feeling in a document. This is also useful for company to determine whether they are viewed positively or negatively by public.The challenge of sentiment analysis is automatically determining whether a text is positive or negative in tone. In our work, we focus on news articles, which use a more neutral vocabulary, as compared to the emotionally charged vocabulary of opinions such as editorials, reviews, and blogs. News analysis is now routinely used by both buy-side and sell-side in market. The main task identified for news opinion mining consists of extracting sentences from online published news articles that mention company news, and identifying positive, negative and neutral sentiment that exists in that text and further summarizing the article polarity. A large number of companies use news analysis to help them to make better business decisions so in our project we are doing sentiment analysis on news article related to company.

Keywords :- machine learning, natural language processing, news analysis, opinion mining, sentiment analysis

I. INTRODUCTION

The sentiment analysis used in Wide range of applications in business and public policy. Sentimental analysis is now being used from specific product marketing to antisocial behavior recognition. The advances in Facebook ,twitter ,YouTube and other micro blogging and social networking sites have not only contributed change to the social sites but have fundamentally changed the way we use these sites and how we share our feelings, our views with the wider audience.

Businesses and organizations have always been concerned about how they are perceived by the public. This concern results from a variety of motivations, including marketing and public relations. Before the era of Internet, the only way for an organization to track its reputation in the media was to hire someone for the specific task of reading newspapers and manually compiling lists of positive, negative and neutral references to the organization. Alternatively, it could undertake expensive surveys of uncertain validity. Today, many newspapers are published online. Some of them publish dedicated online editions, while others publish the pages of their print edition in PDF or similar formats. In addition to newspapers, there are a wide range of opinionated articles posted online in blogs and other social media. This opens up the possibility of automatically detecting positive or negative mentions of an organization in articles published online, thereby dramatically reducing the effort required to collect this type of information. To this end, organizations are becoming increasingly interested in acquiring fine-grained sentiment analysis from news articles.Fine-grained sentiment analysis is an extremely challenging problem because of the variety of ways in which opinions can be expressed. News articles present an even greater challenge, as they usually avoid overt indicators of attitudes. However, despite their apparent neutrality, news articles can still bear a polarity if they describe events that are objectively positive or negative. Many techniques used for sentiment analysis involve naïve approaches based on spotting certain keywords which reveal the author or speaker's emotions. This project presents an opinion-mining engine we have built which performs fine-grained sentiment analysis to classify sentences as positive, negative or neutral.

II. RELATED WORK

The most relevant work is the work done by Simon Fong, Yan Zhuang, Jinyan Li [1] This work presents various Machine Learning (ML) approaches and algorithm comparisons for classification of texts and for doing sentiment analysis efficiently. The text is classified based on three classes positive, negative and neutral classes. This work suggests that it is efficient to use naïve bayes classifier for the pupose of sentiment analysis. As it gives better accuracy as compared to other classifiers used for sentiment analysis. Other classifier used for comparison includes maximum entropy, decision tree, winnow, c4.5 classifiers.

The work given in [2] gives tasks addressed in our work is Semantic parser. The semantic parser provides method for extracting concepts from sentence. This task includes subdividing sentence into verbs, nouns, prepositions, pairs of nouns, adjective and noun pairs. And these are extracted as candidate concepts. This work addresses the fine grained sentiment analysis.

Fine grained sentiment analysis is made commercial viable. [3]In this work opinion mining task is focused. This work proposes a Tweets Sentiment Analysis Model (TSAM) that can capture social interests and people's opinions for a specific social event. This work used Australian federal elections 2010 event as an example. Study of opinions sentiments and emotions expressed in text is sentiment analysis stated by [3]. This works provides working of feature extraction tasks in sentiment analysis. It gives idea that instead of using all the words for sentiment analysis use only those words which carries some opinion. This work explains that building a lexicon based sentiment analysis intelligent system is beneficial Work [3] gives different methods for improving accuracy of classifiers such as naïve bayes for sentiment analysis. They uses negation handling, n-grams, feature selection by mutual information result to improve efficiency. They focuses on generalizes method for number of text categorization problem and improving accuracy and speed.

III. PROPOSED SYSTEM

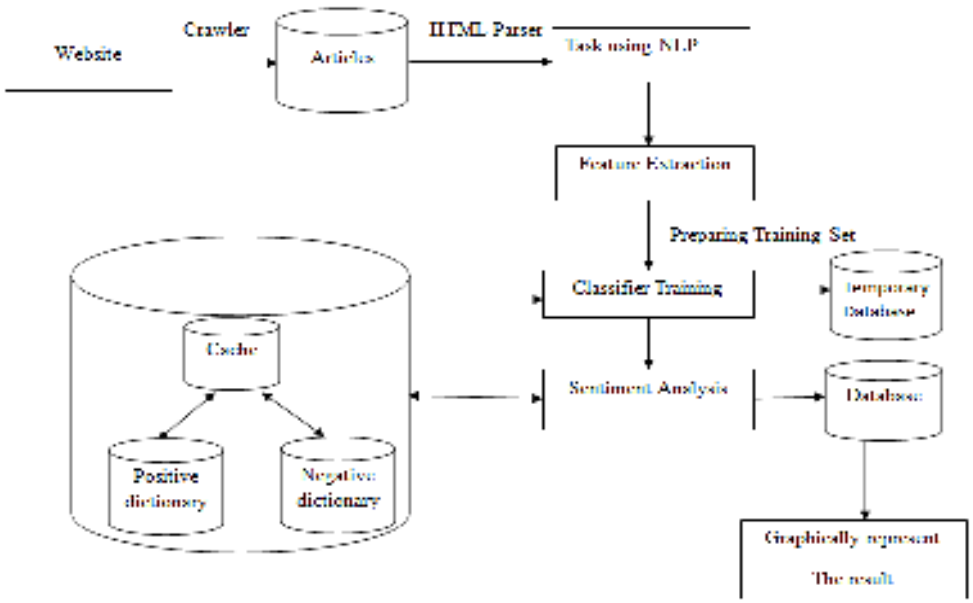

Fig 1.Block Diagram

The first module comprises of two tasks. In the first task, the news articles are downloaded from a website using a web crawler. These articles are in the HTML format. In the second task desired text is extracted from HTML article page. This task can be done using the HTML Parser. The HTML parser selects the desired content from HTML documents and creates a temporary text file.

In the second module data preprocessing steps are performed. The second module is based on the Natural Language Processing (NLP) operations. Once the temporary text file is created, it is subjected to the NLP operations such as Sentence detection, Tokenization, removing punctuations, Parts of speech tagging. These tasks will be done using the WEKA tool. This module gives candidate keywords and combinations of words which will be further useful for determining sentiments of the article.

In third module text classification task is performed. The candidates keywords generated in previous module are taken as input for this task. This candidate keyword is compared with the words in positive dictionary if match found then word is collected in positive class. If word not found in positive dictionary then it will be match with negative dictionary on success word is collected in negative class. This task will be performed using naïve bayes classifier. The information about sentiment is conveyed by adjectives or more specifically by certain combinations of adjectives with other parts of speech. This task of sentiment analysis is performed in this module.

In this module the graphical result is created using positive, negative and neutral count. The graphical result shows sentiment of the corresponding news article. From this sentiment it is determined whether the article is positive, negative or neutral.

A Naive bayes classifier is a simple probabilistic classifier model based on the bayes rule along with a strong independence assumption. The Naïve Bayes model includes a simplifying conditional independence assumption. That is given a class (positive or negative, neutral), the words are conditionally independent of each other.[4] This assumption does not affect the accuracy in text classification by much but makes really fast classification algorithms applicable for the problem.

If the classifier encounters a word that has not been seen in the training set, the probability of both the classes would become zero and there won't be anything to compare between. This problem can be solved by laplacian smoothing Usually, k is chosen as 1. This way, there is equal probability for the new word to be in either class. Since Bernoulli Naïve bayes is used, the total number of words in a class is computed differently. For the purpose of this calculation, each document is reduced to a set of unique words with no duplicates.

Negation handling was one of the factors that contributed significantly to the accuracy of our classifier. A major problem faced during the task of sentiment classification is that of handling negations. Since we are using each word as feature, the word "good" in the phrase "not good" will be contributing to positive sentiment rather that negative sentiment as the presence of "not" before it is not taken into account. To solve this problem [4] devised a simple algorithm for handling negations using state variables and bootstrapping .Generally, information about sentiment is conveyed by adjectives or more specifically by certain combinations of adjectives with other parts of speech. This information can be captured by adding features like consecutive pairs of words (bigrams), or even triplets of words (trigrams).

IV. CONCLUSION

A Thus in this work we have tried to put forth a new methodology sentiment analysis. As the input data source comprises of authenticated news articles, the output yield will be reliable. The algorithms used not only give better results than the other alternatives but also reduce the time required for processing. The results obtained hence, will be more expeditious as well as optimized, due to the use of the fast and accurate naïve bayes classifier, which will guarantee user satisfaction.

REFERENCES

[1] Simon Fong, Yan Zhuang, Jinyan Li, Richard Khoury," Sentiment Anlaysis of Online News using MALLET", 2013 International Symposium on Computational and Business Intelligence ,24-26 Aug 2013, pp 301-304.

[3] Xujuan Zhou, Xiaohui Tao, Jianming Yong, Zhenyu Yang , "Sentiment Analysis on Tweets for Social Events".Proceedings of the 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design. 27-29 June 2013, pp 557562.

[4] Vivek Narayanan1, Ishan Arora2, ArjunBhatia,"Fast and accurate sentiment classification using an enhanced Naive Bayes model".

[5] A.S.M Shihavuddin, Mir NahidulAmbia, Mir Mohammad NazmulArdin, "Prediction of Stock Price analyzing the online financial news using Naive Bayes classifier and local economic trends". 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), 20-22 Aug. 2010, pp V4-22

[6] Seyed-Ali Bahrainian, Andreas Dengel "Sentiment Analysis and Summarization of Twitter Data". 2013 IEEE 16th International Conference on Computational Science and Engineering, 3-5 Dec. 2013

[7] KiranShriniwasDoddi, Dr.Mrs. Y. V. Haribhakta2, Dr.ParagKulkarni " Sentiment Classification of News Articles", KiranShriniwasDoddi et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, pp4621-4623

[8] Wenxin XIONG, Jiajin XU, Maocheng LIANG "An Architecture for Automatic Opinion Classification in Western Online News",IEEE Workshop on Electronics, Computer and Applications, 8-9 May 2014,pp 717 – 721

[9] Esuli,Andrea ,Sebastiani, Fabrizio, " Determining the Semantic Orientation of Terms Through Gloss Classification"In Proceedings of CIKM-05 the ACM SI GIR Conference on Information and Knowledge Management,5 November 2005,pp617

[10] Kamps J ,Marx M,Mokken R J , et aL, "Using WordNet to measure semantic orientation of adjectives". InProceedings of the 4th International Conference on Language Resources and EvMvation,Lisbon,LREC, 2004,ppl115-1118.