SENTIMENT ANALYSIS OF NEWS ARTICLES USING MACHINE LEARNING APPROACH

[1]UBALE SWATI, [2]CHILEKAR PRANALI, [3]SONKAMBLE PRAGATI

[1,2,3]Department of Computer Engg, JSPM's Rajarshi Shahu College of Engineering, Tathawade, Pune
E-mail: swatiubale92@gmail.com[1], chilekarpranali7@gmail.com[2], ashsonkamble@gmail.com[3]

Abstract- Determining the attitude of a writer with Respect to some topic or the overall feeling in a document is basic aim of doing sentiment analysis. News analysis can be used to plot the firm's behavior over time and thus yield important strategic insights about firms. Sentiment analysis is also useful in social media monitoring to automatically characterize the overall feeling or mood of consumers as reflected in social media toward a specific brand or company and determine whether they are viewed positively or negatively. In our work, we focus on news articles. News analysis and news sentiment calculations are now routinely used by both buy-side and sell-side in market surveillance and compliance. The main tasks identified for news opinion mining consists of extracting sentences from online published news articles that mentions company news, and identifying positive and negative sentiment that exist in that article and further summarizing the article polarity. A large number of companies use news analysis to help them 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, Opinion Mining, News Analysis.

I. INTRODUCTION

Wide range of applications in business and public policy uses sentiment analysis. Sentimental analysis is now being used from specific product marketing to antisocial behavior recognition.

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, 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. 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 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 there apparent neutrality, news articles can still bear 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. We use naïve 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 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 purpose 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 Semantic parser. The semantic parser provides method for extracting concepts from sentence. This task includes subdividing sentence into or splitting of sentences. 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 specific social events. 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 use negation handling, n-grams, and Feature selection by mutual information result to improve efficiency. They focuses on generalizes method for number of text categorization problem and improving.

III. SENTIMENT ANALYSIS FRAMEWORK AND TECHNIQUE

The News Sentiment Analysis automatically analyses news articles. It can identify the positive, negative or neutral opinions and measure intensity of positive/negative opinions in regard to an organization. The conceptual framework of the News Sentiment Analysis consists of four modules:

- Crawling and extraction module, Crawl HTML files from specified URL.HTML Parser that extracts desired text from HTML files.
- Data preprocessing and Feature extraction module that perform Natural Language Processing (NLP) operations and extracts the opinionated words from each sentence.
- Sentiment identification, scoring and classifier training module that associates expressed opinions with each entity in each sentence level. Text is classified as positive, negative or neutral class. Sentiment aggregation and scoring calculates the sentiment scores for each entity. Sentiment analysis is performed.
- Sentiment aggregation module that gives Graphical result which is created using positive, negative and neutral polarity count.

Fig. 1 illustrates the News Sentiment Analysis framework. The details of each module are discussed in the following sections.

A. Crawling

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.

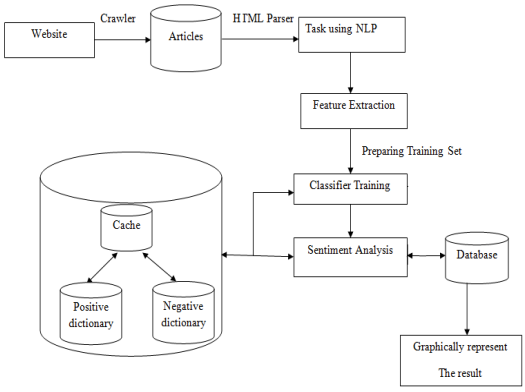

Fig.1. Framework of sentiment analysis.

B. Feature extraction

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.

C. Classifier Training

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.

a) Naïve Bayes Classifier

A Naive bayes classifier is a simple probabilistic classifier model based on the bayes rule along with a strong 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 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. Bernoulli Naïve bayes is also used in this technique for handling duplication .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).

The above described techniques are described below.

- The maximum likehood probability of word belonging to a particular class is given by the expression:

$$P(x_i|c) = \frac{(\text{Count of } x_i \text{ in documents of class } c)}{(\text{Total number of words in documents of class } c)} \quad \ldots (1)$$

- The frequency counts of the words are stored in hash tables during the training phase.
- According to the Bayes Rule, the probability of a particular document belonging to class c_i is given by,

$$P(c_i/d) = P(d/c_i) * P(c_i)/P(d) \quad (2)$$

$$P(c_i/d) = (\prod P(x_i/c_j) * P(c_j) / P(d) \quad (3)$$

We generally estimate $P(c_j)/P(d)$ using m-estimates:

$$P(c_j)/P(d) = nc + mp/ n + m \quad (4)$$

Where:
n = the number of training examples for which $c = c_j$
n_c = number of examples for which $p = c_j$ and $c = c_i$
p = a priori estimate for $P(c_j)/P(d)$
m = the equivalent sample size

Here the x_i s is the individual words of the document. The classifier outputs the class with the maximum posterior probability.

- Laplacian Smoothing

$$P(x_i|c_j) = (\text{Count}(x_i) + k) / ((k+1) * (\text{No of words in class } c_j)) \quad (5)$$

- Mutual Information

$$P(X|Y) = \sum_{y \in Y} \sum_{x \in X} P(x, y) \log\left(\frac{P(x,y)}{P(x)P(y)}\right) \quad (6)$$

D. Result

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.

CONCLUSION

Thus in this work we have tried to present 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 and cost effective methodology will be provided.

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

[2] Prashant Raina, "Sentiment Analysis in News Articles Using Sentic Computing", IEEE 13th International Conference on Data Mining Workshops, 2013, 7-10 Dec.pp 959 – 962.

[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 Narayanan, Ishan Arora, Arjun Bhatia, "Fast And Accurate sentiment classification using an Enhanced Naive Bayes model". A.S.M

[5] Nahidul Ambia, Mir Mohammad Nazmul Ardin, "Prediction of Stock Price analyzing the online Financial News using Naive Bayes classifier and and local economic trends"., 3rd International Conference on Advanced Computer Theory an Engineering (ICACTE) 2010, 20-22 Aug. 2010, pp V4-22

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

[7] Kiran Shriniwas Doddi, Dr. Y. V.Haribhakta, Dr. Parag Kulkarni ,"Sentiment Classification of News Articles",(IJCSIT) International Journal of Computer Science And Information Technologies, Vol.5 (3), 2014, pp 4621-4623

[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 SIGIR Conference on Information and Knowledge Management, 5 November 2005, pp617

[10] Kamps J, Marx M, Mokken R J, "Using WordNet To measure semantic orientation of adjectives",. In Proceedings of the 4th International Conference on Language Resources 2004, ppl115-1118.

★★★