JOURNAL OF INFORMATION, KNOWLEDGE AND RESEARCH IN INFORMATION TECHNOLOGY

THESAURUS CONSTRUCTION OF POLITICAL OPINIONS

[1]MITESH GALA, [2]KUNAL BHATT, [3]JYOTI DESHMUKH, [4]SUNIL DAS, [4]AMIT SINGH

Department of Information Technology, Shah & Anchor Kutchhi Engineering college, University Of Mumbai, Mumbai, Maharashtra, India.

[1] miteshgala1992@yahoo.com; [2]kunalb93@hotmail.com; [3] jyoti.sakec@gmail.com; [4]dassunil8494@gmail.com; [5] amits7844@gmail.com

ABSTRACT : *For the last few decades, public opinion has become an important aspect for every single thing. Be it marketing for a new product or to check the popularity of a celebrity. In this competitive world, it is necessary to know what the common masses think about. Same is the importance of public opinion in the field of politics, even more important. Politicians are completely dependent on the public to stay in power, thus it makes it important for them to know what the public thinks. Even the concept of exit polls is based on sentiment analysis, where one can judge the attitude of the speaker towards any topic. So sentiment analysis is an important aspect in political domain. This report focuses on the development of the Maximum Entropy classifier which aims at determining the positive or negative sentiment of a speaker via his/her comment or review.*

KEY WORDS : Reviews, Politics, Maximum Entropy, Classification, Polarity, Sentiment Analysis

1. INTRODUCTION

Sentiment analysis (also known as opinion mining) in simple man's term is to determine whether the speaker is happy or angry (sad) regarding a topic when he/she gave or posted that comment or review. It not just aims at finding the positivity or negativity, but the focus is also determine the weight factor. This is achieved at word level, then at the sentence level and then finally the sentiment score of the entire comment or review.

This is implemented in multiple modules. The slang words, special characters and stop words are removed in the pre-processing module. The next module focuses on calculation of the empirical weight distribution of the feature words. This is an important step as this is done to normalize the word distribution over the entire data set, which could result in biasness while calculating the polarity and weight. Then comes the classification module, where Maximum Entropy algorithm is used. This is the step where the feature words are compared with Standard English positive and negative words. Then is the step where the proper nouns and other irrelevant words are eliminated. Next, words that satisfy the threshold condition are extracted. This is accomplished by comparing the words with SentiwordNet 3.0. In the finally step, the classifier sums up all the weights for positive and negative words to give a sentiment score for the respective sentence and then the same is done to the entire file in the dataset. Thus the final output is the positive, negative and neutral aspects for the sentence. Lastly, the thesaurus of the classified words is created and displayed.

2. RELATED WORK

Enric Junqué de Fortuny et al [1] presents a survey on the 2011 Belgium elections. Text mining was performed on 68,000 related on-line news articles published in 2011 in Flemish newspapers. These articles were analysed by a custom-built expert system. The results of the text mining analyses showed interesting differences in media coverage and votes for several political parties and politicians. With opinion mining, they were able to automatically detect the sentiment of each article, thereby allowing to visualise how the tone of reporting evolved throughout the year, on a party, politician and newspaper level. The suggested framework introduces a generic text mining approach to analyse media coverage on political issues, including a set of methodological guidelines, evaluation metrics, as well as open source opinion mining tools. Since all analyses are based on automated text mining algorithms, an objective overview of the manner of reporting is provided. The analysis shows peaks of positive and negative sentiments during key moments in the negotiation process.

Alexander Pak et al [2] represent a survey regarding the computational techniques, models and algorithms for mining opinions from unstructured reviews. This paper is divided into 3 subtasks: subjectivity, polarity classification, opinion target extraction, opinion source identification and opinion summarization. The paper first describes what is Opinion Mining (OM) and various applications of it in day to day life. This paper also gives detailed description of various opinion mining disciplines of Natural Language Processing (NLP), text mining, and web mining. Subjectivity & polarity classification is an important

aspect in OM. This research is based on objective classification. It implies appraisal taxonomy. It further classifies the words as positive, negative or neutral. It also shows the calculation of distance between the words to determine their weights. Next task employed was to identify the opinion targets i.e. feature selection, and then the source of the opinions, i.e. to determine the audience or the targeted group from whom the opinions are to be taken. The final task was to conclude whether the document under review was positive, negative or neutral depending upon the weights calculated earlier.

Jayashri Khairnar et al [3] explained about automatic text classification in text mining is a critical technique to manage huge collections of documents. However, most existing document classification algorithms are easily affected by ambiguous terms. The ability to disambiguate for a classifier is thus as important as the ability to classify accurately. In this paper, we propose a novel classification framework based on fuzzy formal concept analysis to conceptualize documents into a more abstract form of concepts, and use these as the training examples to alleviate the arbitrary outcomes caused by ambiguous terms. The proposed model is evaluated on a benchmark test bed and two opinion polarity datasets. The experimental results indicate superior performance in all datasets. Applying concept analysis to opinion polarity classification is a leading endeavor in the disambiguation of Web 2.0 contents, and the approach presented in this paper offers significant improvements on current methods. The results of the proposed model reveal its ability to decrease the sensitivity to noise, as well as its adaptability in cross domain applications.

Bo Pang et al [4] gave a brief idea about micro blogging which is a very common mode of communication among Internet users. Micro blogs are real time content published by people and this content is generally laden with personal opinions about a variety of aspects in everyday life. This makes micro blogs a rich source of data for opinion mining. Here, we use a corpus from the popular micro blogging website, Twitter. We consider micro blogs from the period before the presidential elections in the United States of America in 2012 to analyze the collective sentiment of the micro bloggers, against and in favour of each presidential candidate. We classify the micro blogs to positive and negative opinion classes and we use machine learning classification techniques achieve this. We mainly try to classify the tweets into political opinions against and in favor of the presidential candidates, Barack Obama and Mitt Romney. We used various feature selection techniques and classification algorithms. The best results were obtained by using n-gram features with support vector machines (SVM) classifier. The Twitter corpus used is manually annotated for sentiments and we use this as a gold standard for evaluation of precision, recall and f-score of our classification.

S Chandrakala et al [5] discussed about quality of the interpretation of the sentiment in the online buzz in the social media and the online news can determine the predictability of financial markets and cause huge gains or losses. That is why a number of researchers have turned their full attention to the different aspects of this problem lately. However, there is no well-rounded theoretical and technical framework for approaching the problem to the best of our knowledge. We believe the existing lack of such clarity on the topic is due to its interdisciplinary nature that involves at its core both behavioral-economic topics as well as artificial intelligence. We dive deeper into the interdisciplinary nature and contribute to the formation of a clear frame of discussion. We review the related works that are about market prediction based on online text-mining and produce a picture of the generic components that they all have. We, furthermore, compare each system with the rest and identify their main differentiating factors. Our comparative analysis of the systems expands onto the theoretical and technical foundations behind each. This work should help the research community to structure this emerging field and identify the exact aspects which require further research and are of special significance.

3. PROPOSED SYSTEM

The proposed system is implemented in various modules. The system retrieves the data set from the stored directory, applies pre-processing on that dataset. It then classifies it, finds polarity and weights for negative, positive and neutral words and lastly sums up the respective weights to calculate the final polarity and weight of the files in the dataset.

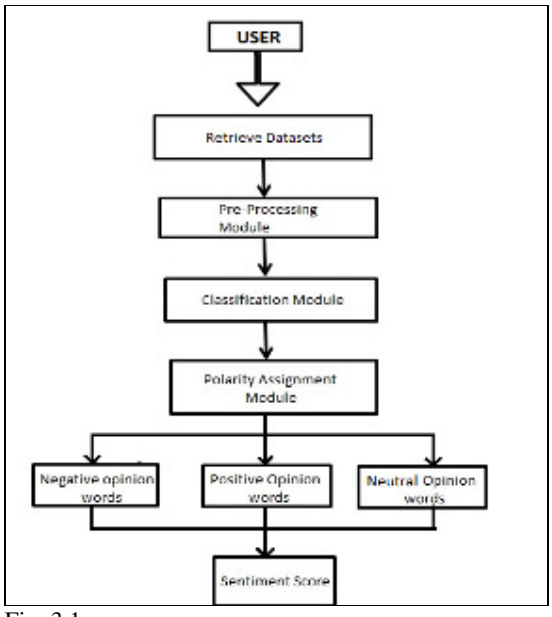

Fig. 3.1

3.1 Retrieving Dataset

This is the module in which the program retrieves the text files one at a time from the dataset folder. The program also allows users to enter reviews dynamically. These reviews will be also stored as text files and retrieved when the classification module is invoked. This output of this module represents all files that were either entered dynamically or were statically present in the program dataset.

3.2 Pre Processing Module

The opinion words which are required for classification will be retrieved. It is important that all stop words are removed and only the feature words are selected. Here the data set will be operated upon as one file at a time. Not only is the removal of stop words important, but slang words and special characters removal is also necessary. Each file will be compared with standard English stop words, slang words and special characters in order to filter them out. An important point to be noted here is that this module will result in some sort of word disambiguation, since words like "not", "very" that are also stop words, would cause the meaning of the conjoint word to change form positive to negative and vice versa. But it is observed that in the political domain, word disambiguation is not a prominent problem. Though it slightly affects the accuracy at the sentence level, the overall accuracy of the classifier is not severely affected.

3.3 Classification Module

This is the core of the project. Before actually moving on to the classification, it is necessary to calculate the empirical distribution of feature words. This is simply the distribution of all the pro-processed words in their respective files and also the normalization of these words over the entire dataset or corpus. This is important because occurrence of a negative or a positive word multiple times in a single file may cause the classifier to produce bias results, thus normalization is performed.

Once the empirical distribution is calculated, classification is performed on the empirical words. In the classification process, the program compares each word with a set of English positive and negative words. Those words that fail to be classified as either positive or negative are labeled as neutral. Once classification is performed, classified words compare the words with SentiwordNet 3.0, which is standard list of sentiment words along with their weight. We consider this weight as the threshold value, and the empirical words that fail to meet this threshold are eliminated. This implies that at every step, the program filters out the words that may not affect the classifier accuracy greatly.

3.4 Sentiment Score

Sentiment score is what predicts whether the file is positive, negative or neutral. This is implemented at two levels. Firstly, the sentiment score of each sentence in a given file is calculated. To accomplish this, the classifier sums up all the positive values, the negative values, and the neutral values of the feature words of that particular sentence and displays the weights for each. This way, a glossary for every file is created stating which sentences are positive, which are negative and those which are positive, along with their weights. The classifier accuracy for sentences was calculated to be 84.34%.This is further summed up, i.e. the weights and their polarities of each sentence in a particular file are summed up to show the positive, negative and neutral aspect of each single file as a whole. For a tagged dataset of 35 files, the classifier has an accuracy of 77.14%.

3.5 Thesaurus Construction

Once the sentiment score and the polarity have been determined for all the empirical words, the final step is to display the thesaurus. This thesaurus is constructed using SentiwordNet 3.0. Since SentiwordNet 3.0 has a synonym or meaning for every word in it, the thesaurus constructed will contain the polarity of the text file, the empirical words that were identified in each of these text file, their sentiment score calculated by the classifier and the meaning of those words derived from SentiwordNet 3.0. This will allow the user to understand how and why the prediction of the file was positive, negative or neutral and provide a comparative analysis of the various files that were taken as a dataset.

4. ALGORITHM

Input:

- *D_p : Dataset of political domain*

Output:

- *Text sentiment score(Lexicon with polarity value)*
- *Dataset sentiment score*
- *Polarity*

Steps:

[1] Read Dataset D_p.

[2] Perform pre-processing i.e. slang words, special characters and stop word removal on entire dataset

[3] Calculate Emperical distribution

[4] Classification:

[4.1] For all text files in dataset D_p calculate

f_{ij} = frequency of term i in the document j.

Also, calculate

d_{fi} = document frequency of term i

id_{fi} = inverse document frequency of term i

$id_{fi}/log\ (N/df_i)$ Where, N=sample size

[4.2] Normalize the term frequency across the entire corpus

$$tf_{ij} = f_{ij}/\max\{f_{ij}\}$$

$$\vartheta = 1/M\ log \sum idf_i$$

$$And\ M = idf_i$$

$$Initialize\ \lambda = 0$$

$$And\ \lambda_{i+1} = \lambda_i + \vartheta$$

$$P_\lambda(y/x) = 1/Z_\lambda(x)\ exp(\sum \lambda_i f_i(x,y))$$

Where,

$Z_\lambda(x)$ = *the normalizing factor to ensure exponential probability given by*

$$Z_\lambda(x) = \sum exp(\sum \lambda_i f_i(x, y))$$

Where

λ_i= *parameter associated with constraint fi to be estimated*

[5] Using SentiWordNet, classify words as positive, negative or neutral.

In the text file in D_p

[6] For all emperical words

$Pos\_weight = \sum(positive\ words)$

$Neg\_weight = \sum(negative\ words)$

$Neu\_weight = \sum(neutral\ words)$

[7] If

$Pos\_weight \geq Neg\_weight$

$Sentence = positive$

Else if

$Pos\_weight \leq Neg\_weight$

$Sentence = negative$

Else

$Sentence = neutral$

[8] Repeat step [7] at file level

5. RESULTS AND DISCUSSIONS

The proposed project gives a detailed analysis of the political dataset or reviews or opinions under study. The system has successfully managed to predict the sentiment polarity i.e. whether the review is positive, negative or neutral. Not only does the system tell you that, but it also illustrates the final number of words on which classification was performed. This greatly helps to determine the accuracy of the classifier. Using the classifier program on a dataset of 35 text files, total of 111 words were classified and the classification accuracy of 84.34% was achieved at the sentence level. At the file level, using the same tagged dataset for of 35 text files, an accuracy of 77.14% was achieved.

A few important points were noted during the implementation of this project. Firstly, word disambiguation problem was encountered. It was found that due to word disambiguation, a reply given by a Member of Parliament, India in the Rajya Sabha was classified as negative, which was otherwise positive. Secondly, inspirational and persuasive speeches that include greater number of lower weight negative words, out numbered the high weight positive words, leading to predict the speech as negative. Also, during the classification process, the empirical words that fall below the required threshold have also had an impact on the overall accuracy. It was also observed that the neutral weight weights associated with a file was because of the words not being completely classified as negative or positive, but because they tend to depict the positive aspect of the review.

5. Conclusions

Thesaurus Construction of Political opinions is aimed at classifying political opinions, reviews and comments as positive, negative or neutral using the Maximum Entropy algorithm. This implementation has proved to be highly accurate at both the sentence level and the file level having accuracies of 84.34% and 77.14% respectively.

The proposed method is preferred over the Naïve Bayesian approach because it does not separate the data into various classes, nor are any special labels required for it. And the biggest advantage that it has over the Naïve Bayes method is that it does not consider different classes, in this case the various empirical words to be independent of each other, specially within the same text, hence Maximum Entropy approach explains the relation between among words (or classes) in a more effective manner. Though, word disambiguation will remain a major problem, it can be corrected at the pre-processing stage.

6. REFERENCES:

[1] Enric Junqué de Fortuny, John Lafferty, Andrew McCallum,“ Using Maximum Entropy for Text Classification” IJCAI-99 Workshop on Machine Learning for Information Filtering, 1999, Pages 61-67 – Max Entropy.

[2] Alexander Pak and Patrick Paroubek, “Twitter as a corpus for Sentiment Analysis and Opinion Mining”, Proceedings of the Seventh conference on International Language Resources and Evaluation, pp. 1320-1326, 2010.

[3] Jayashri Khairnar and Mayura Kinikar, “Machine Learning Algorithms for Opinion Mining and Sentiment Classification”, International Journal of Scientific and Research Publications, Volume 3, Issue 6,pp. ISSN 2250-3153 June 2013.

[4]Bo Pang and Lillian Lee. 2008. “Opinion Mining and Sentiment Analysis”. Found. Trends Inf. Retr. 2, 1-2 (January 2008), 1-135

[5] S CHANDRAKALA and C SINDHU, “OPINION MINING AND SENTIMENT CLASSIFICATION: A SURVEY”, Department of Computer Science and Engineering, Velammal Engineering College, India .

[6] JEEVANANDAM JOTHEESWARAN and DR. Y. S. KUMARASWAMY,” Opinion Mining Using Decision Tree Based Feature Selection Through Manhattan Hierarchical Cluster Measure”, Journal of Theoretical and Applied Information Technology, 10th December 2013. Vol. 58 No.1, ISSN: 1992-8645.

[7] Tamara Martín-Wanton, Aurora Pons-Porrata, Andrés Montoyo-Guijarro, Alexandra Balahur,” OPINION POLARITY DETECTION Using Word Sense Disambiguation to Determine the Polarity of Opinions”, Center for Pattern Recognition and Data Mining, Universidad de Oriente, Patricio Lumumba s/n, Santiago de Cuba, Cuba, Department of Software and Computing Systems, University of Alicante, Alicante, España.

[8] Muhammad Zubair Asghar, RahmanUllah, Bashir Ahmad, Aurangzeb Khan, Shakeel Ahmad and Irfan Ullah Nawaz," POLITICAL MINER: OPINION EXTRACTION FROM USER GENERATED POLITICAL REVIEWS", Sci.Int(Lahore),26(1),385-389,2014, ISSN 1013-5316

ABOUT THE AUTHORS

Mitesh Gala is currently doing his BE from Shah & Anchor Kutchhi Eng. College, Mumbai. His areas of interest include Java and Android programming. He also has a keen interest in web development.

Kunal Bhatt is currently doing his BE from Shah & Anchor Kutchhi Eng. College, Mumbai. His areas of interest include database technologies, software testing and web development.

Jyoti Deshmukh is currently working as Assistant Professor in Information Technology Dept. of SAKEC, Mumbai. She is working toward the PhD degree in Computer Science. Her research interest include opinion mining, information retrieval and natural language processing.

Sunil Das is currently doing his BE from Shah & Anchor Kutchhi Eng. College, Mumbai. His areas of interest include various web development technologies like HTML, .NET, ASP, XML.

Amit Singh is currently doing his BE from Shah & Anchor Kutchhi Eng. College, Mumbai. His area of interest includes software development.