International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in; editorijoer@gmail.com

Vol.5., Issue.2, 2017
March-April

RESEARCH ARTICLE

ISSN: 2321-7758

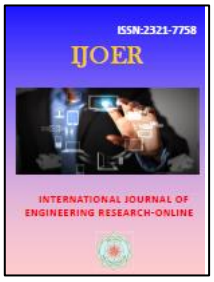
BIG DATA IMPLEMENTATION OF UNSTRUCTURED DATA ANALYTICS OF SOCIAL NETWORK REVIEWS USING SENTIMENT ANALYSIS & SVM

KEERTHANA M[1], JEYA MOHAN H[2]

[1]UG Student, [2]Assistant Professor

Department of Computer Science and Engineering, Alpha College of Engineering, Chennai, T.N, India

[1]keerthana.muthumani@gmail.com;[2]jeyamohan2519@gmail.com



ABSTRACT

In the Exiting system, We notice that these sentiment dictionaries have numerous inaccuracies. We could not able to principally categorize the Opinion Results. Sentiment based analysis is the major key in categorizing the user's Feedback. We are using FSM & EEM Algorithm for the Word processing process. In this paper the modification Process, Twitter like Application is created and users Tweets are processed. We are implementing Big Data in this Project. Users Tweets are the input to the Big Data HDFS System. Data are stored in the Data Nodes. Index is maintained in the Name Node. Tweets are clustered and classified based on Keywords extracted, stemmed and tokenized. Each tokens are analysed using Sentiment Analysis and classified as Positive and Negative opinion for Tweets. Map & Reduce is implemented.

Key Words: Big Data, Data mining, Sentiment Analysis, SVM (support vector machine).

I. Introduction

The Big Data opinion mining is becoming an important tool to improve efficiency and quality in organizations, and its importance is going to increase in the coming years. It is the important aspect for capturing public opinion about product preferences, marketing campaigns, political movements, social events and company strategies. In recent times, research activities in the areas of Opinion, Sentiments and/or Emotions in natural language texts and other social media are gaining momentum based on subjectivity or objectivity analysis. The reason may be the huge amount of available text data in the social Web in the forms of news, reviews, blogs, chats and even twitter . In particular, Twitter contains various forms of information, such as video links, image links, and text data. In this study, we focus on text data and we study methods for extraction of sentiment information from big-data text. The opinion words are extracted using the resulting frequent features, and semantic orientations of the opinion words are identified with the help of WordNet. The system then finds those infrequent features. The part of speech tagging from natural language processing is used to find opinion features. Thus, text summary of opinions is generated . Summarization work is truly dependent on the features and hence is far from the automatic summarization work in the field of NLP. This proposes a method by utilizing the adjective synonym set and antonym set in WorldNet to predict the semantic orientations of adjectives. The ”semantic orientation" is a measure of subjectivity and opinion in text. It deals with the actual text element. It transforms it into a format that the machine can use.

In the Sentiment analysis is treated as a classification task as it classifies the orientation of a

text into either positive or negative. The general approach of determining the overall orientation (i.e., positive or negative) of a sentence/ document is by analysis of the orientations of the individual words. Sentiment dictionaries are utilized to facilitate the summarization. There are numerous works that, given a sentiment lexicon, analyse the structure of a sentence / document to infer its orientation, the holder of an opinion, the sentiment of the opinion, etc.

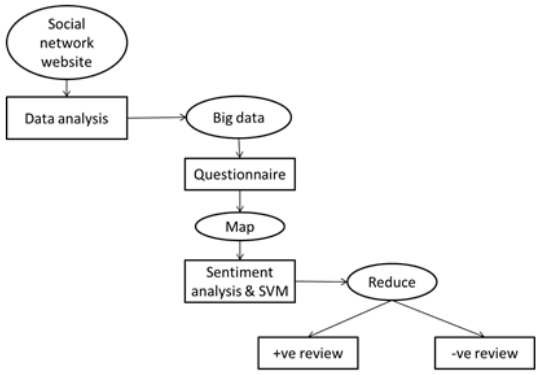


Fig 1: Architecture Diagram

II. Related work

A. *The Role of Text Pre-processing in Sentiment Analysis*

Sentiment analysis in reviews is the process of exploring product reviews on the internet to determine the overall opinion or feeling about a product. Reviews represent the so called user-generated content, and this is of growing attention and a rich resource for marketing teams, sociologists and psychologists and others who might be concerned with opinions, views, public mood and general or personal attitudes. The huge number of reviews are on the web content represents the current form of users feedback. It is hard for human companies to get the latest trends and summarise the state or general opinions about products due to the big diversity and size of social media data, and this creates the need of automated and real time opinion extraction and mining. Deciding about the sentiment of opinion is a challenging problem due to the subjectivity factor which is essentially what people think. Sentiment analysis is treated as a classification task as it classifies the orientation of a text into either positive or negative. Machine learning is one of the widely used approaches towards sentiment classification in addition to lexicon based methods and linguistic methods. It has been claimed that these techniques do not perform as well in sentiment classification as they do in topic categorisation due to the nature of an opinionated text which requires more understanding of the text while the occurrence of some keywords could be the key for an accurate classification.

B. *Classification of Sentimental Reviews Using Machine Learning Techniques*

Sentiment Analysis is the most prominent branch of natural language processing. It deals with the text classification in order to determine the intention of the author of the text. The intention can be of admiration (positive) or criticism (Negative) type. The dataset considered for training and testing of model in this work is labelled based on polarity movie dataset and a comparison with results available in existing literature has been made for critical examination. an attempt has been made to classify sentiment analysis for movie reviews using machine learning techniques. Two different algorithms namely Naive Bayes (NB) and Support Vector Machine (SVM) are implemented. These two algorithms have also been implemented earlier by different researchers and results of all versions of implementation have been compared. It is observed that SVM classifier outperforms every other classifier in predicting the sentiment of a review. In this study, only two different classifiers have been implemented. In future, other similar classification strategies under supervised learning methodology like maximum entropy classifier, stochastic gradient classifier, K nearest neighbour and others can be considered to implement and a comparison of results can be presented with SVM classifier.

C. *Map Reduce Functions to Analyze Sentiment Information from Social Big Data*

Opinion mining, which extracts meaningful opinion information from large amounts of social multimedia data, has recently arisen as a research area. In particular, opinion mining has been used to understand the true meaning and intent of social networking site users. It requires efficient techniques to collect a large amount of social multimedia data and extract meaningful information from them. Therefore, in this paper, we propose a method to extract sentiment information from

various types of unstructured social media text data from social networks by using a parallel Hadoop Distributed File System (HDFS) to save social multimedia data and using Map Reduce functions for sentiment analysis. The proposed method has stably performed data gathering and data loading and maintained stable load balancing of memory and CPU resource during data processing by the HDFS system. The

Proposed Map Reduce functions have effectively performed sentiment analysis in the experiments. Finally, the sentiment analysis results of the proposed system are very close to those of manual processes.

D. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts

Sentiment analysis seeks to identify the view point(s) underlying a text span; an example application is classifying a movie review as .thumbs up or thumbs down. To determine this sentiment polarity, We propose a novel machine-learning method that applies text-categorization techniques to just the subjective portions of the document. Extracting these portions can be implemented using efficient techniques for finding *minimum cuts in graphs* the minimum-cut framework results in the development of efficient algorithms for sentiment analysis. Utilizing contextual information via this framework can lead to statistically significant improvement in polarity-classification accuracy. Directions for future research include developing parameter selection techniques, incorporating other sources of contextual cues besides sentence proximity, and investigating other means for modelling such information.

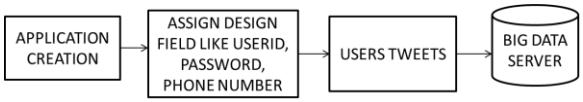
E. Determining the Sentiment of Opinions

Identifying sentiments (the affective parts of opinions) is a challenging problem. We present a system that, given a topic, automatically finds the people who hold opinions about that topic and the sentiment of each opinion. The system contains a module for determining word sentiment and another for combining sentiments within a sentence. We plan to extend our work to more difficult cases such as sentences with weak-opinion-bearing words or sentences with multiple opinions about a topic. To improve identification of the Holder, we plan to use a parser to associate regions more reliably with holders. We plan to explore other learning techniques, such as decision lists or SVM None the less, as the experiments show, encouraging results can be obtained even with relatively simple models and only a small amount of manual seeding effort. Improves identification of the holder. Work more on difficult cases such as sentence with weak opinion bearing on topic.

III. MODULE FRAMEWORKS

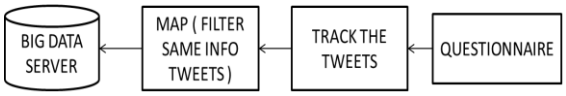
A. TWITTER PROCESS

In this module we will create an application to tweet with our friends. For creating an Application, we will be using Advanced Java Concepts like JSP and Servlets. While creating the application, we'll assign the design fields like Username, Password, Phone and other information. Once the created the user is allowed to enter the data. Also the server will store the data and allow the user to enter in to the chat application. The User will enter the tweets through this application.



B. MAPPING OF TWITTER ANALYSIS

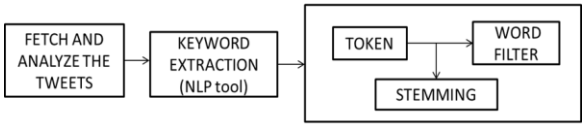
In this module our centralized server will track all the user information from different tweets and can effectively filter the users who have used same information in different tweets. In our real situation user posts different topics like politics, corruption etc. Based on that server are mapping and analysis user tweets. We try to gather all these users and collect their tweets via centralized server.



C. REDUCE – CLUSTER ANALYSIS

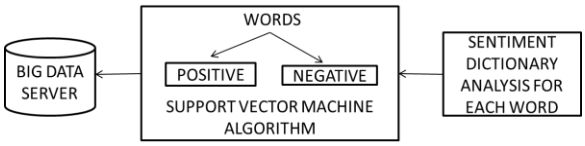
In this module, output of mapping part is the input of Reduce part. The Server will analyse the Tweets between the Users and the extract the Keywords using NLP. The NLP will the extracts the Keywords and filters the other words using the Stemming Algorithm. By using the Stemming algorithm we can filter the unwanted words in the

chat so that we can calculate the extracted words counts.



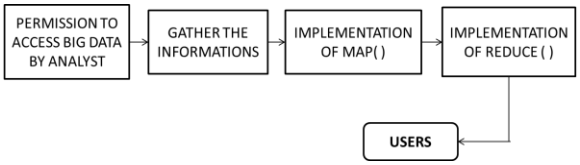
D. SVM AND SENTIMENTAL ANALYSIS

In this module we introduce the sentiment analysis that analyze the words of tweets comparing with sentiment dictionary and classify it using SVM to gather and analyze user tweets of corresponding issues in social network with gets input among the people in all around the world. In this we not getting feedback using graphical mode, we introduce a SVM method so that the feedback in text given to the machine will understand the text and classify as positive and negative with the trained datas.



E. DATA GATHERING AND CLUSTERING OPINION POLL

In this module we implement big data, in this big data we will have lot or vast amount of data that may wanted or unwanted information in simple the information in the big data are unstructured. So in this module the insurance server is going allow permission to access the server by the big data analyst .The big data analyst get the all the information which mention above and extract the information by the technique of map reducing formation to get useful information.



IV. METHODOLOGY/ALGORITHM

A. Natural Language Processing

Natural language processing techniques plays important role to get accurate sentiment analysis. NLP techniques like Bag of words, Hidden markov model, part of speech (POS), N-gram algorithms, large sentiment lexicon acquisition and parsing techniques are used to express opinion for document level, sentences level and aspect level. The opinion words are extracted using the resulting frequent features, and semantic orientations of the opinion words are identified with the help of WordNet. The system then finds those infrequent features. The part of speech tagging from natural language processing is used to find opinion features. Thus, text summary of opinions is generated . Summarization work is truly dependent on the features and hence is far from the automatic summarization work in the field of NLP. This proposes a method by utilizing the adjective synonym set and antonym set in WordNet to predict the semantic orientations of adjectives. The lexicon-based approach involves calculating sentiment polarity for a review using the semantic orientation of words or sentences in the review. The —semantic orientation" is a measure of subjectivity and opinion in text. It deals with the actual text element. It transforms it into a format that the machine can use.

B. STEMMING PROCESS

Stemming is a process for removing the commoner morphological and inflexional endings from words in English. Its main use is as part of a term normalisation process that is usually done when setting up Information Retrieval systems.

Successor variety stemmers:

The successor varieties for a given word have been derived; this information must be used to segment the word.

-cutoff method, peak and plateau method, complete word method, entropy method.

To illustrate the use of successor variety stemming, consider the example below where the task is to determine the stem of the word READABLE.

Test Word: READABLE

Corpus: ABLE, APE, BEATABLE, FIXABLE, READ, READABLE, READING, READS, RED, ROPE, RIPE.

Prefix	Successor Variety	Letters
R	3	E,I,O
RE	2	A,D
REA	1	D
READ	3	A,I,S
READA	1	B
READAB	1	L
READABL	1	E
READABLE	1	BLANK

Using the complete word segmentation method, the test word "READABLE" will be segmented into "READ" and "ABLE," since READ appears as a word in the corpus. The peak and plateau method would give the same result.

After a word has been segmented, the segment to be used as the stem must be selected following rule:

if (first segment occurs in <= 12 words in corpus)
first segment is stem
else (second segment is stem)

the successor variety stemming process has three parts: (1) determine the successor varieties for a word, (2) use this information to segment the word using one of the methods above, and (3) select one of the segments as the stem.

C. SUPPORT VECTOR MACHINE

SVM can be applied to various optimization problems such as regression; the classic problem is that of data classification. The basic idea is shown in figure 1. The data points are identified as being positive or negative, and the problem is to find a hyper-plane that separates the data points by a maximal margin

$$u = \vec{w} \cdot \vec{x} + b$$

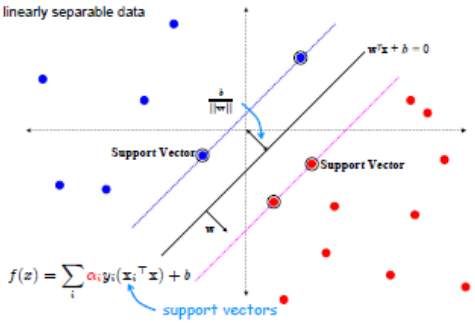


Figure 2 : Data Classification

The above figure only shows the 2-dimensional case where the data points are linearly separable. The mathematics of the problem to be solved is the following:

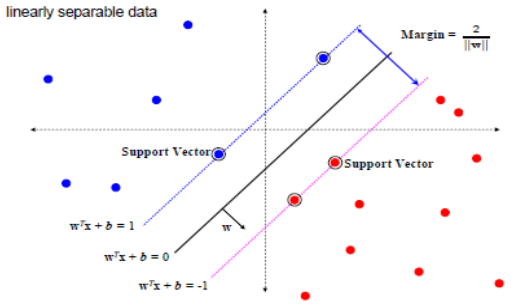
$$s.t \quad y_i = +1 \Rightarrow \vec{w} \cdot \vec{x}_i + b \geq +1$$
$$y_i = -1 \Rightarrow \vec{w} \cdot \vec{x}_i - b \leq -1$$

$$s.t \; y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1, \quad \forall i$$

The identification of the each data point x_i is y_i, which can take a value of +1 or -1 (representing positive or negative respectively). The solution hyper-plane is the following:

$$u = \vec{w} \cdot \vec{x} + b$$

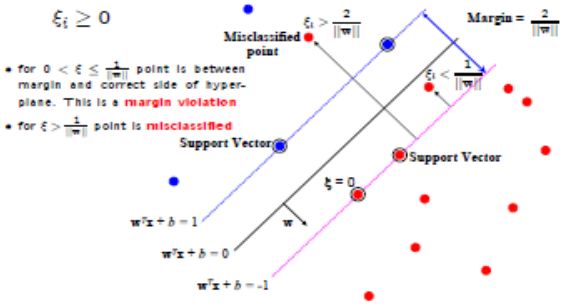
The scalar b is also termed the bias, w is weight vector.



When the hyperplane misclassifies then:

$$\frac{w.x + b}{\|w\|} \geq w$$

Maximize the bias and minimize the weight vector.



A standard method to solve this problem is to apply the theory of Lagrange to convert it to a dual Lagrangian problem. The dual problem is the following:

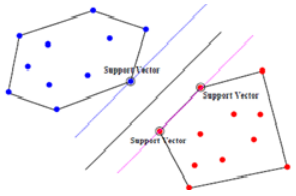
$$\min_{\alpha} \Psi(\vec{\alpha}) = \min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j (\vec{x}_i \cdot \vec{x}_j) \alpha_i \alpha_j - \sum_{i=1}^{N} \alpha_i$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

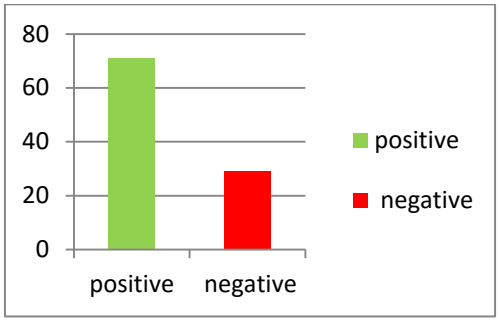
$$C \geq \alpha_i \geq 0, \quad \forall i$$

The variables α_i are the Lagrangian multipliers for corresponding data point x_i.

$$u(x) = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b$$



Corresponding issue result as



V. CONCLUSION

In this project, we have to implement and analyze the user tweets from social network like twitter, face book etc. Every user posts and tweets about any issue like politics, education department, cinema, environment, social incident, etcs are the inputs in our project. The Map & Reduce technique is used to extract the exact keyword to analyses the tweets using sentiment analyze technique with sentiment dictionary. SVM algorithm is used to classify the positive, negative commands based on corresponding issue. So classification of positive, negative reviews of the corresponding issue get easier. As this is from big data People can easily identify whether the corresponding issue going on is positive or negative in short duration from large data along with high effectiveness. SVM is used to get most accuracy level of classification for the process of opinion polling and to categorize principally.

REFERENCES

[1]. Emma Haddi, Xiaohui Liu, Yong Shi, "The Role of Text Pre-processing in Sentiment Analysis", in Int Conf on Info Tech, 2013 , Procedia Comp Science 17 (2013) 26 – 32.

[2]. Abinash Tripathy, Ankit Agrawal, Santanu Kumar Rath, "Classification of Sentimental Reviews Using Machine LearningTechniques", Int .conf Procedia Computer Science 57 (2015) 821 – 829.

[3]. Ilkyu Ha, Bonghyun Back, and Byoungchul Ahn, "MapReduce Functions to Analyze Sentiment Informationfrom Social Big Data", in IJDSN Article ID 417502,Volume 2015.

[4]. B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in Proc. 42nd Annu. Meeting Assoc. Comput. Linguistics, 2004, pp. 271–278.

[5]. M. Kim and E. Hovy, "Determining the sentiment of opinions," in Proc. 20th Int. Conf. Comput. Linguistics, 2004, pp. 1367–1373

[6]. C. Danescu-N.-M., G. Kossinets, J. Kleinberg, and L. Lee, "How opinions are received by online communities: A case study on amazon.com helpfulness votes," in Proc. 18th Int. Conf. World Wide Web, 2009, pp. 141–150.

[7]. H. Takamura, T. Inui, and M. Okumura, "Extracting semantic orientations of words using spin model," in Proc. 43rd Annu. Meeting Assoc. Comput. Linguistics, 2005, pp. 133–140.

[8]. E. Breck, Y. Choi, and C. Cardie, "Identifying expressions of opinion in context," in Proc. 20th Int. Joint Conf. Artif. Intell., 2007, pp. 2683–2688.

[9]. X. Ding and B. Liu, "Resolving object and attribute coreference in opinion mining," in Proc. 23rd Int. Conf. Comput. Linguistics, 2010, pp. 268–276.

[10]. A. L. Maas, R. E. Daly, P. Pham, D. Huang, A. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, 2011, pp. 142–150.

[11]. P. Stone, D. Dunphy, M. Smith, and J. Ogilvie, The General Inquirer: A Computer Approach to Content Analysis. Cambridge, MA, USA: MIT Press, 1996.

[12]. A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," Project Report, Stanford University, 2009.