International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017),e-ISSN:2455-2585
International Conference on
Recent Explorations in Science, Engineering And Technology (ICRESET'19)
Volume-5, Special Issue-March, 2019.

Twitter Based Sentiment Analysis

Anitharajam.M[1], Atchaya.S[2], Dheekshitha.D[3], Suganthi.A[4]

Department of computer Science and Engineering,KGISL Institute of Technology,Coimbatore,Tamilnadu,India

*Abstract---**Sentiment analysis over Twitter offer organisations a fast and effective way to monitor the publics' feelings towards their brand, business, directors, etc. A wide range of features and methods for training sentiment classifiers for Twitter datasets have been researched in recent years with varying results. In this paper, we introduce a novel approach of adding semantics as additional features into the training set for sentiment analysis. For each extracted entity (e.g. iPhone) from tweets, we add its semantic concept (e.g. "Apple product") as an additional feature, and measure the correlation of the representative concept with negative/positive sentiment. The social web has made enormous amounts of information available to users globally at just the click of a button. Consumers often tend to rely on such text, especially those in the form of opinions or experiences regarding a particular product which makes it essential that this information should be available in a systematic manner. Sentiment analysis studies these opinions. This paper explains different methods for sentiment analysis and showcases an efficient methodology. It also highlights the importance of Naïve Bayes classifier over other classification algorithms.***

Keywords---Sentiment Analysis, Machine learning, Naive Baye's, Support Vector

I. INTRODUCTION

Social Computing is an innovative and growing computing pattern for the analysis and modelling of social activities taking place on various platforms. It is used to produce logical and interactive applications to derive efficient results . The wide availability of social media sites provides individuals to share their sentiment or opinions about a particular event, product or issue. Mining of such informal and homogeneous data is highly useful to draw conclusions in various fields. Though, the highly unstructured format of the opinion data available on web makes the mining process challenging . Textual information present on web is majorly classified into either of the two categories: fact data and sentiment data. Fact data are the objective terminologies concerning different entities, issues or events. Whereas sentiment data are the subjective terms, that define individual's opinions or beliefs for a particular entity, product or event. Sentiment analysis is the process of recognizing and classifying different sentiments conveyed online by the individuals to derive the writer's approach towards a specific product, topic or event is positive, negative. Sentiment analysis is carried out at different levels ranging from coarse level to fine level. The coarse level sentiment analysis determines the sentiment of the whole manuscript or document. The fine level sentiment analysis, whereas focuses on the attributes. Sentiment analysis of Twitter data is carried out on sentence level which comes in between coarse level and fine level. In the sentiment analysis process, the sentiments present in the text are of two types: Direct and Comparative. The direct sentiments in text are independent from other objects in the same sentence .In this paper; we present a sentiment analysis process for Twitter data. Twitter is a micro-blogging site that is rapidly growing in terms of number of users. Moreover, Tweets are mostly public and limited to 140 characters that simplify the identification of emotions in text . Though, the abundance of data, use of short forms, timing of different posts, and diversity of language make the sentiment analysis process difficult for Twitter data.

II. PROBLEMS IN EXSISTING SOLUTIONS

As per the thorough literature survey, the major identified techniques are as follows: (A) Lexicon Based Model –It employs frequent and explicit product features extraction involving Syntax Tree Based Classification-Design Syntactic Patterns.(B) Word Alignment Model (Unsupervised)-It concerns with Word Co-occurrence Frequencies and Position of Words. (C) Word Alignment Model (Semi-supervised) - It involves analysis of Formal and Informal Text Separately Based on the above techniques, these are the various models that have been identified and some of the related models are discussed below. It is possible to combine some features from Word Alignment Model and Lexicon Based Model to design a new semi-supervised lexicon based model so that it is possible to use lexical databases like WordNet, SentiWordNet and Attempto Controlled English Lexicon [ACE].

Among these lexical databases WordNet groups English words into sets of synonyms called synsets. SentiWordNet processes unstructured information and extracts meaningful numeric indices from the text and aims to provide an extension for WordNet such that all synsets can be associated with a value concerning the negative, positive or objective connotation. ACE provides deep classification of parts of speech but it is better to use ACE along with WordNet to increase recognition rate of lexemes.

III. PROPOSED SYSTEM

- Proposed system will gives you the freedom to choose the data of any topic.
- It gives you the impact the results and statistics will have on the respective field.
- Proposed system allows retrieval of data based on the query entered by the user.
- Proposed system will provide accurate feature selection.

IV. METHODOLOGY FOR SENTIMENT ANALYSIS

The sentiment analysis of Twitter data is an emerging field that needs much more attention. Fig. 1 shows the steps to carry out the process of sentiment analysis on Twitter data. Firstly, the collected Twitter data is pre-processed to perform the data cleaning. Secondly, the important features are extracted from the clean text, applying any of the feature selection methods. Thirdly, the portion of the data is manually labeled as positive or negative Tweets to prepare a training set. Finally, the extracted features and the labeled training set are provided as an input to the built classifier to classify the remaining data i.e. Test set. Each of the processing steps is discussed thoroughly in the following sub-sections.

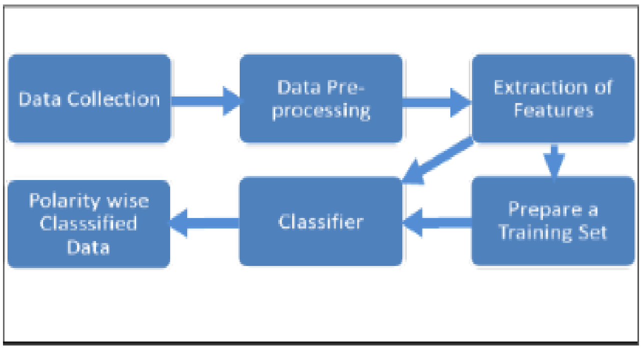

Fig.1 The process of sentiment analysis on Twitter data

A. Twitter Growth Rate Statistics

Approximately 6,000 tweets are tweeted on Twitter on per second basis. It resembles to 350,000 tweets sent per minute and 500 million tweets per day. That makes it around 200 billion tweets per year. In Twitter's history, the number of Tweets increased from 5,000 tweets per day in 2007 to 500,000,000 tweets per day in 2013,that is approximately a six orders of magnitude. At the intermediate stages it has the statistics of 300,000 tweets per day in 2008,2.5 million tweets per day in 2009, 35 million tweets per day in 2010, 200 million tweets per day in 2011.And 340 million tweets per day six years after the emergence of Twitter i.e. On March 21, 2012. This statistics conclude the use of Twitter for our research.

B. Feature Extraction

The pre-processed dataset has various discrete properties. In feature extraction methods, we extract different aspects such as adjectives, verbs and nouns and later these aspects are identified as positive or negative to detect the polarity of the whole sentence. Followings are the widely used Feature Extraction methods.

- **Terms Frequency and Term Presence**: These features denote individual and distinct words and their occurrence counts.
- **Negative Phrases:** The presence of negative words can change the meaning or orientation of the opinion. So it is evident to take negative word orientation in account.
- **Parts Of Speech (POS)**: Finding nouns, verbs, adjectives etc, as they are significant gauges of opinion

C. Sentiment Classification Techniques

Machine learning techniques are further classified into supervised and unsupervised techniques. To carry out sentiment analysis, typically the supervised machine learning techniques are used as we are dealing with subjective data. Supervised machine learning techniques highly depend on training data which are already labeled data unlike in the case of unsupervised machine learning techniques. Based on the provided training data, the classifier will classify the rest data i.e. Test data. A large number of supervised machine learning algorithms such as Logistic Regression, Naïve Bayes, Decision Tree, Support Vector Machine (SVM), Random Forest, Maximum Entropy, and Bayesian Network are used for sentiment analysis . Choice of an appropriate algorithm for selected data and domain is a crucial step.

D. Naïve Bayes (NB) Approach

Naïve Bayes classifier is a simple probabilistic classifier that uses the concept of mixture models to perform classification. The mixture model relies on the assumption that each of the predefined classes is one of the components of the mixture itself. The components of the mixture model denote the probability of belongingness of any term to the particular component. Thus, they are also known as generative classifiers. Naïve Bayes classifier is a probabilistic classifier that uses the concept of Bayes Theorem and finds maximum prospect of probability of any word fitting to a particular given or predefined class. The probability P is defined as follows:

$$P\,(Xi \mid c) = \frac{Count\ of\ Xi\ in\ document\ of\ class\ c}{Total\ no\ of\ words\ in\ document\ of\ class\ c}$$

Where Xi is a given term and c is a predefined class label. During the training phase, the incidence counts of the words are collected and stored in the hash tables. NB approach suffers from an assumption that the features are independent in the feature space. As per the definition of probability, the document d is classified into class c using following equation:

$$c *= argmax\, P\,(c \mid d)$$

E. Support Vector Machine (SVM)

Support vector machine (SVM) solves the traditional text categorization problem effectively; generally outperforming Naïve Bayes as it supports the concept of maximum margin. The main principle of SVMs is to determine a linear separator that separates different classes in the search space with maximum distance i.e. with maximum margin . If we represent the tweet using t, the hyper plane using h, and sentiment of the tweet. Classes using a set Cj € {l, -1} into which the tweet has to be classified, the solution is written as follows equivalent to the

$$\overline{h} = \sum i\ ai\ Ci\ \overrightarrow{tj}\,, \qquad ai \geq 0$$

The idea of SVM is to determine a boundary or boundaries that separate distinct clusters or groups of data. SVM performs this task constructing a set of points and separating those points using mathematical formulas.

V. RELATED WORK

Current research focuses on sentiment analysis of information gathered from social networking websites like Twitter, Facebook, and MySpace to conclude viewers' response to a particular social event or issue. Sentiment analysis has endless applications like forecasting market movement based on news, blogs and social media. Currently, sentiment analysis is a very beneficial approach for hefty applications like 'Smart Cities'. These applications use methods based on document level and sentence level classification which use purely supervised or unsupervised classification algorithms. These algorithms are advanced by Fuzzy Formal Concept, Genetic Algorithms or Neural Networks by making them semi-supervised. Research also focused on sentiment analysis with networking to give a degree of parallelism. It focused on online accrue utility scheduling algorithm which gave them high speed on multiple processors. But this made the system much more complex. Research was also focused on Twitter sentiment analysis for security-related information gathering using normalized lexicon based sentiment analysis. While it provided a positive outcome, a universal dataset was not used. Current online product recommendation applications are comparing parameters like price, ratings and special offers on the product on different e-commerce websites and are not focusing on customers' personal experience by analyzing their reviews. Hence there is a need to develop a comprehensive application based on sentiment analysis which will give more importance to customer reviews.

VI. CONCLUSION

The task of sentiment analysis, especially in the domain of micro-blogging, is still in the developing stage and far from complete. So, we propose a couple of ideas which we feel our worth exploring in the future and may result in further improved performance.

Right now we have worked with only the very simplest unigram models; we can improve those models by adding extra information like closeness of the word in a negation word. We could specify the window prior to the word (a window could for example be of 2 or 3 words) under consideration and the effect of negation may be incorporated into the model if it lies within that window. The closer the negation word is to the unigram word whose prior polarity is to be calculated, the more it should effect the polarity. For example if the negation is right next to the word, it may simply reverse the polarity of that word and farther the negation is from the word the more minimized its effect should be. Apart from this, we are currently only focusing on unigrams and the effect of bigrams and trigrams may be explored.

REFERENCES

[1]Sentiment Analysis of product reviews E-commerce recommendation D.mali 2016

[2] Mrs. Sayantani Ghosh, Mr. Sudipta Roy, and Prof. Samir K., "A tutorial review on Text Mining Algorithms", International Journal of Advanced Research in Computer and Communication Engineering, Pages: 223-233, Vol. 1, Issue 4, June 2012

[3] Christopher D. Manning, Prabhakar Raghavan, HinrichSch¨utze, "Introduction to Information Retrieval", ISBN-13 978-0-511-41405-3, 2013

[4] HuLi, Yong Shi"WordNet based lexicon model for CNL" 2009 IEEE proceeding at 2009 International Conference on Test and Measurement

[5] Kang Liu, Liheng Xu, and Jun Zhao "Co-Extracting Opinion Targets and Opinion Words from Online Reviews Based on the Word Alignment Model"2014 IEEE proceeding at IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING

[6] V.K. Singh, R. Piryani, A. Uddin P. Waila, Marisha "Sentiment Analysis of Textual Reviews ,Evaluating Machine Learning, Unsupervised and SentiWord Net Approaches" proceeding in IEEE 2013 5th International Conference on Knowledge and Smart Technology (KST)

[7] Mrs. Sayantani Ghosh, Mr. sudipta Roy, Prof. Samir K. Bandyopadhyay "A tutorial review on Text Mining Algorithm" Proceeding in International Journal of Advanced Research in Computer and Communication Engineering 2012.

[8] Pablo Gamallo,MarcosGarcia"Citius: A Naïve Bayes Strategy for Sentiment Analysis on English Tweets" Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 171–175, Dublin, Ireland, August 23-24 2014.

[9] Ricardo Baeza-Yates, Berthier Ribeiro-Neto,"Modern Information Retrieval", ISBN-13: 978-0201398298, 2013

[10] Federica Bision, Paolo Gastaldo, Chiara Peretti, Rodolfo Zunino and Erik Cambria "Data Intensive Review Mining for Sentiment Classification across Heterogeneous Domains" 2013 IEEE at ACM International Conference on Adavances in Social Networks Analysis and Mining