Survey of Classification of Business Reviews Using Sentiment Analysis

Shilpa A. Shendre[1], Prof. Pramila M Chawan[2]

[1]*M.Tech Student, Dept of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India*
[2]*Associate Professor, Dept of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India*

---***---

ABSTRACT:- *The rapid increase in mountains of unstructured textual data accompanied by the proliferation of tools to analyze them has opened up great opportunities and challenges for text research. The research area of sentiment analysis has gained popularity in the last years. Business developers not only want to know about there product marketing and profit based on the number of sales been done but also want to know about the reviews and thoughts of people using these products. The feedback they receive via social media and other internet services becomes very important to measure the quality of a product they are serving. Sentiment analysis is a domain where the analysis is focused on the extraction of feedback and opinions of the users towards a particular topic from a structured, or unstructured textual data. In this paper, we try to focus our effort on sentiment analysis on Yelp challenge database. We examine the sentiment expression to classify the reviews of the business whether it is positive or negative and perform the feature extraction and use these features for updating and maintenance of the business.*

Key Words: *sentiment analysis; opinion mining; classification; text reviews, Machine learning*

1. INTRODUCTION

Sentiment analysis has become an important research area for understanding people's opinion on a matter by differentiating a huge amount of information. The present era of the Internet has become a huge Cyber Database which hosts the gigantic amount of data which is created and consumed by the users. People across the world share their views about various services or products using social networking sites, blogs or popular reviews sites. The Internet is been growing at an exponential rate giving rise to communicate across the globe in which people express their views on social media such as Facebook, Twitter, Rotten Tomatoes and Foursquare. Opinions which are being expressed in the form of reviews provide a platform for new explorations to find collective reviews of people. One such domain of reviews is the domain of business reviews which affects business people. The feedback from the customer is valuable for companies to analyze their customer's satisfaction and survey the competitors. This is also useful for other people or consumers who want to buy a product or a service prior to making a purchase.

In this paper, we are going to present the results of machine algorithms for classifying reviews using semantic analysis. A large number of customer-generated reviews for businesses and service providers are classified as either positive or negative. We propose a method to automatically classify customer sentiments using only business text review. This helps us to generate the result using feedback without manual intervention. By studying only rating, it is very difficult to judge why the user has rated the product as 1 or 5 stars. However, the text content contains a more quantitative value for analyzing more than rating itself.

In this paper, we are going to mention the preprocessing steps require in order to achieve accuracy in the classification task. There is no previous research available on classifying sentiment of business review using the latest reviews forms yelp dataset. Determining the underlying sentiment of business review is a difficult task taking into account several factors such as the connotation of a word depending on the context, language used, words ambiguity when using words that don't express a particular sentiment or when using sarcasm. We show that a sentiment analysis algorithm built on top of machine learning algorithms such as Naïve Bayes and Linear Support Vector Classification (SVC) has accuracy above 90% business reviews.

1.1 Feature Selection

Mostly the researchers apply standard feature selection in there approach to improve performance with few using more practical approaches. We are focusing completely on feature election to improve sentiment analysis are few. One of them is the famous Pang & Lee, who removed objective sentences on a tested consisting of objective and subjective text trained on SVM. Initially, they found that sentiment classification result is actually slow and moderate. They then concluded it was more likely that sentences adjacent to discarded sentences improved classification result over their baseline.

1.2 Information Gain

Another work used sophisticated feature selection and found that using either information gain (IG) or genetic algorithm (GA) results in an improvement inaccuracy. Let D be a dataset of labeled texts. Let pD represent the probability that a random text D is classified as positive. The classification should be fairly simple if the text is majorly biased towards positive or negative instances. On the contrary, if the set is very unevenly distributed with equal likelihood of positive and negative instances, then the task is difficult. The disorder in the set D is calculated by its entropy:

$$H(D) = -p_d \log(p_d) - (1 - p_d)\log(1 - p_d)$$

This can be simplified as the average number of bits I required to communicate the classification of each item in the corpus. It is required to choose relevant features that help us classify the set D. A feature is useful if it helps to lower disorganization in the corpus. On choosing a feature x, the corpus is divided amongst instances where x is 0 and instances where x is 1. Let the subsets be D0 and D1. If both of these sets are relatively organized, then we have minimized disorder. Quantitatively, information gain is calculated by:

$$IG(D,x) = H_2 - \frac{|D_0|}{|D|} H_2(D_0) - \frac{|D_1|}{|D|} H_2(D_1)$$

This is the difference between the entropy in the original dataset D, and the average entropy of the sets D0, D1. The Information Gain Criterion chooses features *x1,.....xk* that maximize IG(D,k). It chooses one feature at a time.

1.3 SentiWordNet

SentiWordNet (SWN) is an extension of WordNet that was developed by Esuli & Se- bastiani, which augments the information in WordNet with sentiment of the words in them. Each synset in SWN comprises of sentiment scores that are positive and negative score along with an objectivity score. The summation of these three scores gives the relative strength of positivity, negativity and objectivity of each synset. These values have been obtained by using many semi-supervised ternary classifiers, with the capability of determining whether a word was positive, negative, or objective. If all the classifiers settled on a result then the highest value are assigned for the analogous score, else the values for the positive, negative and objective scores were proportional to the number of classifiers that assigned the word to every class.

2. LITERATURE SURVEY

Hu et al. perform the classification of a document at the sentence level. Instead of the whole document and feature extract on which views have been expressed, identifying comments words by proposing a technique that uses the WordNet

lexical database. For each feature extracted, the related reviews sentence is stored in positive or negative categories and computes a total count. The features are ranked on the bases of there frequency of the appearance in the reviews. The feature-based summary of the reviews of the product sold online was provided by the authors.

Usually work related to sentiment analysis using machine learning techniques in determining if the overall review is positive or negative movie reviews as data. The writer's used unigram model and Navie Bayes, entropy classification, and SVM to perform the classification and achieve

accuracy upto 80%. They finally concluded that their results outperform the method based on human tagged features.

A system was built by Blair-Goldensohn et al. which automatically summarize sentiment from a set of reviews for a local service such as restaurant or hotel and combine the review sentiment per aspect such as food, service, decor, value etc., Basically they have implemented a custom built lexicon based on WordNet and used a classifier at the sentence level.

To capture word sentiments, the writer used a supervised learning algorithm based on similarities between words which takes into account the rating of previous reviews for capturing the representation of words vectors. On a dataset of about 10 000 reviews from a dataset, the accuracy reported is about 70%. Other work on opinion mining, and in particuSlar in review mining using dataset, focuses on predicting a business' rating based on its only reviews' text for reducing the bias of the users. The authors create a stack of words representation of the most frequent words in all text reviews or most frequent adjectives after POS combined with Linear Regression, Support Vector Regression and Decision Tree Regression. Using 10,000 reviews from Yelp Academic dataset, the Root Mean Square Error (RMSE) is around 0.8[6].

Previously worked depend on sentiment analysis and yelp dataset reviews used to focus on predicting star rating using the text alone. The writer experiment algorithms of machine learning such as Naive Bayes, Perceptron, a Multicast SVM on a sample of more than 10,000 reviews from the dataset. For feature selection and preprocessing techniques such as removing stop words or stemming i.e reducing the words to their root form., they use Bing Liu Opinion Lexicon method.

3. Proposed System

The basic methodology to determine polarity is the one with a lexical approach, where we look at the words comprising the document and apply some algorithms to quantify words with some sentiment score and determine the collective polarity. We have based our computational method on the publically available library SentiWordNet [4].In this work for determining the polarity of the document, we have focused on two areas: 1) Feature Selection and Ranking 2) Classification using Machine Learning techniques. We use the Rotten Tomatoes movie review dataset comprising 8000 reviews. We tend to label the polarity as follows: 0- Strong Negative, 1-Weak Negative, 2- Neutral, 3-week, Positive, 4-Strong Positive. The proposed methodology can be well explained from the below figure.

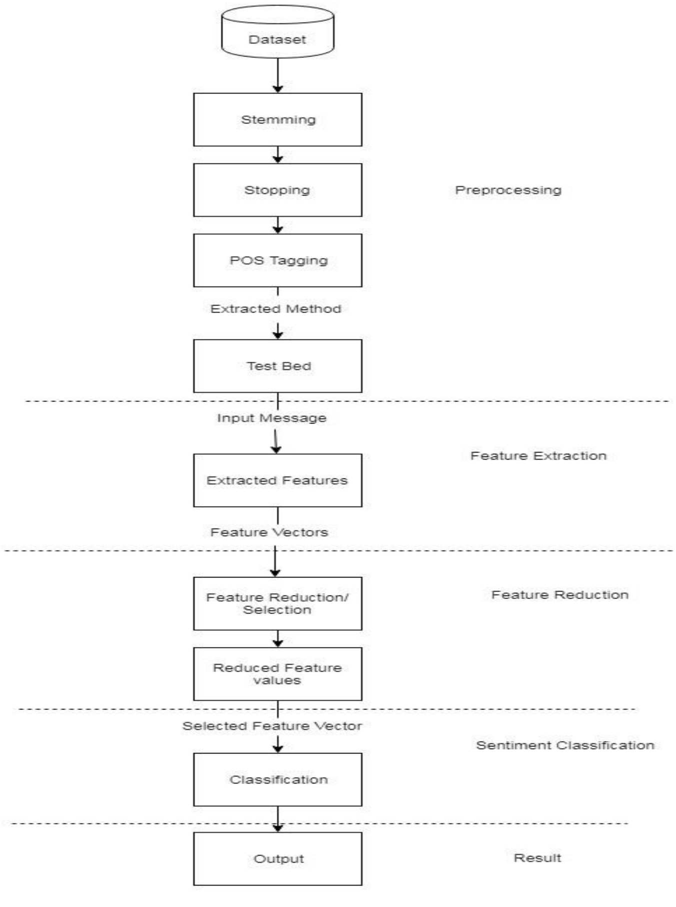

Fig. Proposed Sytem

4. Conclusions

This paper examines the benefits of future extraction methods and classifiers for differentiating the customer reviews using huge dataset. The best classifier is Linear SVC and SGD have given us accurate result upto 94.4% using the first approach proposed in the feature extraction algorithm. In terms of execution, Naive based and Logistic Regression classifiers proved to have a slightly worst result. Sentiment expression, opinion, feelings, and emotion are the tough task in a human analysis. Sentiment analysis ted to perform properly in classifying sentiments for Yelp business reviews by considering star rating given by customers.

References

[1] Kudakwashe Z, Oludayo O., et al. “A Framework for Sentiment Analysis with Opinion Mining of Hotel Reviews.2018”. ICTAS

[2] Kulvinder S.,& Sanjeev D., &Pratibha "Real-time Data Elicitation from Twitter: Evaluation and Depiction Strategies of Tweets Concerned to the Blazing Issues Through Twitter Application " LREC. Vol. 10. 2010.

[3] Rincy J.,& Varghese S. "Prediction of Election Result by Enhanced Sentiment Analysis on Data using Classifier Approach " Computational linguistics 267-307.

[4] Xiaobo Z., &Qingsong Y., "Hotel Reviews Sentiment Analysis Based on Word Vector Clustering " 2017 2nd IEEE International Conference on Computational Intelligence and Applications.

[5] Abdullah A., Sumeet Agrawal., Seon H., Cyrus S "Geo-spatial Multimedia Sentiment Analysis in Disasters” 2017 International Conference on Data Science.

[6] Andreea Salinca. "Business reviews classification using sentiment analysis 17th ISSANA

[7] Wang Z., Qin S., A Sentiment Analysis Method of Chinese Specialized Field Short Commentary” 2017 3rd IEEE International Conference on Computer and Communications

[8] Harpreet K., Veenu M., Nidhi., “A Survey of Sentiment Analysis techniques International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2017) Mining Workshops. IEEE, 2011.

[9] Xian F., Xiaoge L., Feihong D., Xin L., Mian W., “Apply Word Vectors for Sentiment Analysis of APP Reviews” in The 2016 3rd (ICSAI 2016)

[10] Jie Li; Lirong Qiu, “A Sentiment Analysis Method of Short Texts in Microblog” in 2017 IEEE International Conference on Computational Science and Engineering (CSE)

[11] Kudakwashe Z., Oludayo O., “A Framework for Sentiment Analysis with Opinion Mining of Hotel