Sentiment Analysis of a Product based on User Reviews using Random Forests Algorithm

Dr. Shailendra Narayan Singh
Computer Science and Engineering
Amity School of Engineering and Technology
Amity University,Noida, India
Snsingh36@amity.edu

Twinkle Sarraf
Computer Science and Engineering
Amity School of Engineering and Technology
Amity University,Noida, India
Sarraf.twinkle0501@gmail.com

***Abstract*— After many sentiment analysis as well as many types of methods classify the reviews that is based on test data and reviewer's ratings which uses training. , after reading reviews it is seen that star rating of reviewer do not always give a precise measure of his sentiment. This paper primarily focuses on analyzing customer reviews from the e-commerce space. Upon surveying popular e-commerce websites it can be observed that in several instances the product rating given by a customer is not consistent with the product review written by him/her. The problem is made complex by the fact that there is no standard scale to measure the rating that the user gives and the rating of the product are instinctive to the customers' view. In several cases it is seen that a product is rated 4 out of 5. However, the reviews detail that the customer's experience with the product is not favourable. Indeed, text reviews are a true picture of the product. To get rid of this problem, the stated system will give a boolean result i.e. whether the product is good or bad and the user does not need to read all the reviews to analyze the product.**

Keywords used — *Sentiment analysis, product reviews, random forest classifier, bag-of-words*

I. INTRODUCTION

Formally this sentiment analysis has been referred to as a kind of analysis using natural language processing, some sort of computational linguistics as well as text mining. When a n individual makes a decision, it may be the decision or thought influenced by others influence. Moreover, internet provide a forum for this. We can take the example of flipkart.com customer feedback system for the rating of products that they receive from Flipkart and at the same time allows other customers to make a more informed decision by making the ratings available to other customers to review before they make a purchase decision. Almost every business organisation today are in rush to realize that whether individuals like their items and administrations, what do customers consider, what sort of things individuals truly like and don't like, item, benefit which may truly help organisations to settle on choices. These days the greater part of the general population don't purchase things without making some examination of the item over the web, individuals check for the item audits/reviews and after that settle on their choices. Back in the time when organisations required the general population or buyers' conclusions when they need the conduction of opinion surveys that maybe costly and additionally tedious and will require human resource.

This presents challenges which may not be easily addressed by the aid of simple text classification tactics. Thereupon, there is a need to incorporate techniques for classifying opinions into a simplified text classification tool, or to develop systems that will be able to accurately analyze and classify sentiments in text. A kind of contextual mining which helps to identify and extract subjective information or subjective data is called sentiment analysis. This type of extraction helps a business to understand their brand's social sentiment or product or services when they monitor online conversation. This need for analyzing the sentiments has risen in recent years due to the application of sentiment analysis in varied areas such as business intelligence, research, public relations, e- governance and web search Other factors that tends to develop in the increased rate in sentiment analysis are discussed below -

- The rising of methods of machine learning and retrieval of particular data.
- The availability of data sets in the field of machine learning algorithms which must be trained, in general there must be culmination of websites related to review-aggregation.
- The realization of difficulties that are offered by the intelligence and commercial applications.

978-1-7281-2791-0/20/$31.00 ©2020 IEEE

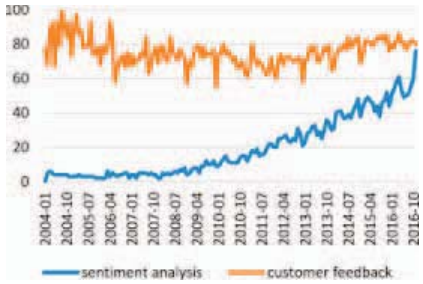


Figure1: Google Trends data showing the relative popularity of search strings "Sentiment analysis" and "Costumer feedback"

II. STRUCTURAL DESIGN OF OPINION MINING

A typical type of sentiment analysis model is given in the figure 1. This model processes the reviews that it takes as input by three important steps: Preparation of data, review Analysis as well as Classification of Sentiment. This model produces output for classification of reviews.

A. Data Preparation

The pre-processing of data and cleaning on the reviews is performed by the data preparation step for the continuous analysis. Some of the frequently used steps of pre-processing contains of contents such as '.',',' etc. and HTML tags as well as removal of irrelevant data from the reviews for sentiment analysis, such reviewers' name and review date[3].

B. Review Analysis

For extracting all the relevant and interesting information like the opinions and analyzing the linguistic characteristics of reviews, the second step of review analysis is performed. This step before extracting suggestions and product characters, processes the opinion by applying different types of tasks which is computational linguistic in nature. The opinion analysis procedure then steps to extract opinion from processed reviews.

C. Sentiment Classification

The basic two types of approaches to classifying reviews are:

1) Machine learning approach
2) SO approach

We will discuss about machine learning approach for sentiment classification in this paper. This type of approach is somewhat similar to classifying top into positive sentiment class and negative sentiment class. The reviews are then stepwise broken down in phases or in the words, represents the review in the form of a document vector defining the opinions which is based on document vectors.

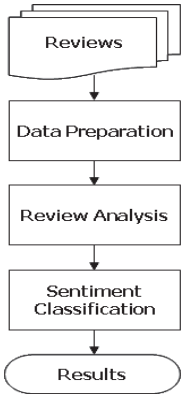


Figure2: A typical sentiment analysis model

III. RELATED WORK

Sentiment Analysis under the topic of micro-blogging is a topic which is recent in the research topics and there can be even more research related to this topic. A large amount of related work related to reviews, documents is done on sentiment analysis of above factors including general phase level sentiment analysis. We may take naive bayes and supporting vector machines being supervised learning machines as the best results but for the supervised approach, but the manual labeling required is very expensive. Approaches that are applied are semi-supervised and unsupervised and there could be many more improvements. Many different researchers who compare their result to the base-line other performances. These proper and formal result comes after these comparison for the selection of best and most efficient features and classification technique. Therefore, the proper performances comes after the comparison.

Among many ways, of the sediment analysis, in this paper we are going to make a focus with the use of machine learning approaches on sentiment analysis.

IV. PROPOSED APPROACH

In this work the live data of certain e-commerce sites will be collected using their respective URLs. The data collected will include customer reviews about the product. The system will crawl the URLs in order to find the opinions. Here we are scrapping the data from internet so that we can have exact opinions as per user requirement.The data crawled from the website will be parsed in order to extract the reviews, which will be subjected to analysis and processing.

This system is different from the existing system in the sense that no pre-existing data sets are used but the live data currently running on the sites i.e. the latest reviews and ratings of the users are analyzed to calculate the final boolean result of the product. This system analyses the text based opinion of the user using natural language processing techniques and puts to use the word model's bag. A popular feature which depicts simplicity as well as good performance is the bag-of-word. The model helps to represent the text with no connection of

the words with each other. This model is one of the popular models and is very useful in the process of sediment analysis and many other researches. The best and simple method to include this model in our classifier is by using uni-grams. A collection of particular word in the text to be classified , where the word once used is not affected by usage of other word .
The Bag of Words model tells about the vocabulary used in the sentences as many times but counted only once as:

Sentence 1: "The bell fell "
Sentence 2: "The audible bell once fell in the well"
This makes our vocabulary as :
{ the, bell, fell, in, audible, the, once, well, ate, and }
The bag of words can be analyzed by the multiple usage of a word occurred in each sentence. For example "the", "bell" "fell"" each appear once.
To limit the limitation of feature vector, a particular size of vocabulary must be chosen. The most frequent 5000 words are used here.
The sentiment analyzer of reviews after cleaning of reviews and implementing Bag of Words model for uses **Random Forest Classifier** for classification. This section focuses on various forests classifier, as well as the impact they make related to accuracy and other features. The first paper made that focuses on the ensemble of decision trees that was composed of multiple tree combination were random forest. Problems like noise or outliners may occour that may affect the result in the single tree classifier, whereas random forest is much robust to noise as well as the provided randomness. Two types of randomness namely bagging and bootstrapping are the main concept of random forest classifier.

Algorithm of Random Forest

Input: No. of trees= B, Training Data = N, Total Features, f = Subset of Features
Output: for input data, bagged class level.

1.Analysing each tree in forest B:
a) Selection of a bootstrap sample S of the size N from training data.
b) creation of tree Tb repeating recursively the following steps:
i. randomly choosing f from F.
ii. Selection of best from F.
iii. Splitting of the node.

2. After creating B trees, the instance of test will be passed to each tree as well as there will be assignment of class label o the basis of majority of votes.

V. IMPLEMENTATION AND RESULT

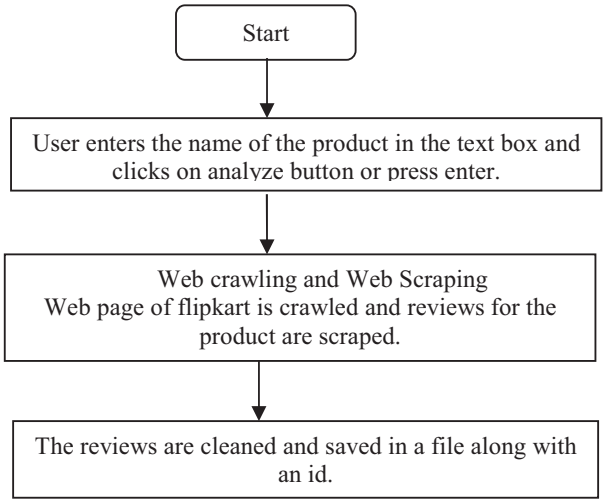
Data in the form of raw reviews have been scraped from Flipkart.com using BeautifulSoup library of python. BeautifulSoup pulls the data of HTML and XML files out. Using your favourite parser to give out the various ways of search, navigate and modify the parse tree.

When the user enters the product's name, the product is goggled and web crawling is done. The URL of the search result page must be known before any code performed. On any search on google, the browser's address bar contains the URL:
https://www.google.com/search?q=SEARCH_THE_TERM_HERE the page is then downloaded after the search. Finally, the web browser module is used to open the browser tabs for that link. Now Flipkart has been appended to the url so as to search for the product on flipkart. After this again beautiful soup is used to crawl the links on the web page. The first link is crawled finding the data-href under the r class. After having crawled the link, next task is to scrap the reviews. The reviews have been scraped by again using beautiful soup. The contents/ reviews are scrapped and stored in comma separated files along with the review id. The tab separated input file is fed to the code which implements Random Forests algorithm. Forests are constructed for training dataset. The constructed forests will be traversed to arrive at the sentiment of each review in input file. If the number of positive sentiments for an input file is >= 50% of the total number of input reviews, then the product is recommended. Otherwise, it is not.

The output is also stored in the form of comma separated file containing id, review and sentiment as columns.

Initially the raw data is loaded. The sentiment analysis is done using Natural language processing and after that vader which is a part of NLTK module can be used . It uses a lexicon of words to find negative and the positive ones. It also analyses the sentiments to determine the sentiment scores. After which vader returns four values for each text:

- A neutrality score
- A positivity score
- A negativity score
- An overall score that summarizes the previous score



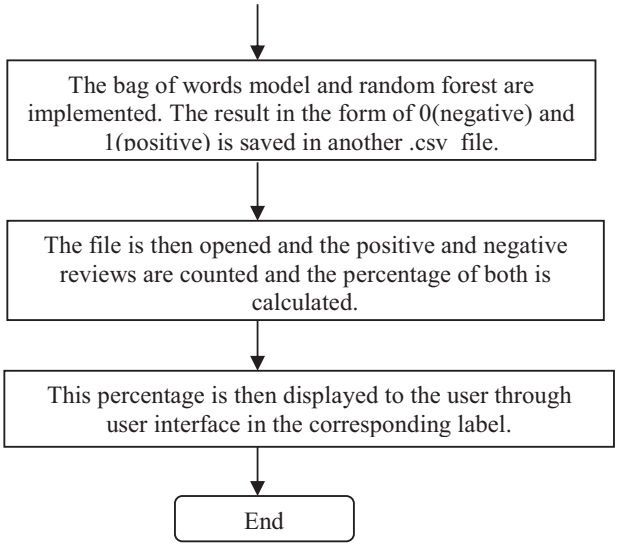


Figure3: Control Flow Diagram of the system

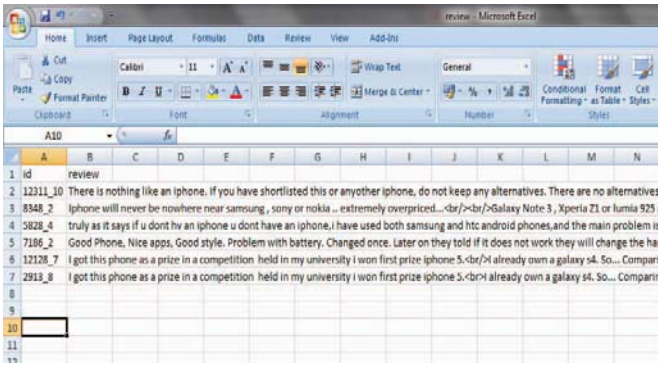


Figure4: File Review.csv where the scraped reviews are saved

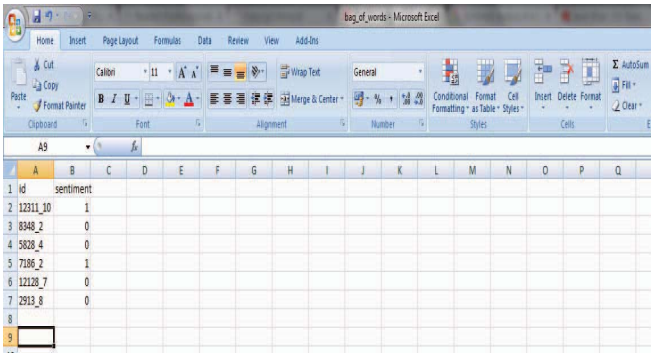


Figure5: File Bag_of_words .csv after implementing Bag-of-words model and random forest algorithm

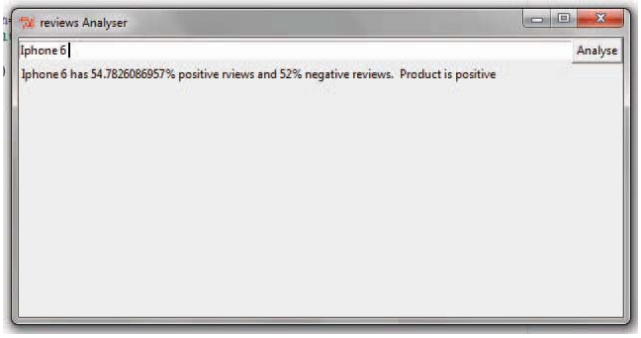


Figure6: User Interface displaying the final result for entered product

V. CONCLUSION

Sentiment analysis contains some sentiments whose classification of text must be dealed with. This paper consists of typical sentiment analysis model which comprises of most importantly three core steps that is named as data preparation, review analysis and sentiment classification as well as It describes representative techniques. Sentiment analysis involves the research in the field of text mining and computational linguistics. It was a good option to attract the significant research attention since last some years.

The issue of rating – review disparity is addressed by this tool. It gives boolean review based on reviews and not on ratings which avoids the problems mentioned earlier. The reviews are extracted from e-commerce site Flipkart.com so that the reviews are accurate.. It gives boolean review of product. Therefore, it will be easier to decide. The count of negative as well as positive reviews are also displayed to show the precision of recommendation. The boolean result whether the product is recommended or not is displayed in user friendly interface.

.

VI. FUTURE ENHANCEMENTS

In the domain of micro-blogging, the working of sentiment analysis is still developing an it is far from complete. We should get some ideas to explore in the future development and performance improvisation. The project can be enhanced to take input as the Url of the product.

The reviews can then be extracted directly from the entered url. The tool currently displays the sentiment based on both product reviews and seller reviews. It can be further enhanced to display sentiment based on product reviews and

seller reviews separately. It can be web hosted with a different database format. It can be extended for some more e-commerce sites. The tool currently analyses reviews only from Flipkart. It can further be enhanced for more than one e-commerce sites like Amazon, eBay etc. so looking for just our topic and focusing on the uni-gram, further exploring bi-grams and tri-grams. As when used with bi-grams, uni-grams usually enhanced I it's performance.

REFERENCES

[1] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner and Isabell M. Welpe. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In Proceedings of AAAI Conference on Weblogs and Social Media (ICWSM), 2010.

[2] Bo Pang and Lillian Lee. (2008). Opinion mining and Sentiment analysis. Foundations and Trends in Information Retrieval Vol. 2, Nos. 1–2. Pages: 1–135

[3] **Stephen** C. F. Chan, Cane W. K. Leung* Sentiment Analysis of Product Reviews

[4] B. Pang , L. Lee, and S. Vaithyanathan, Thumbs up?:sentiment classification using machine learning techniques, Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol.10, 2002, pp. 79-86

[5] Hitesh Parmar, Sanjay Bhandari, Glory Shah (2014, July). Sentiment mining of Movie Reviews using Random Forest with Tuned Hyperparameters. Presented at: International conference on Information Science, Kerala. [Online]

[6] Bing Liu. (2010). Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing, Second Edition. (editors: N. Indurkhya and F. J. Damerau)

[7] L. Breiman, Random forests, Machine Learning, vol. 45. Issue 1, pp. 5-32G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)

[8] Apoorv Agrawal, Boyi Xie, Owen Rambow, Sentiment analysis of Twitter data. Columbia University, New York, NY 10027 USA.[Online]

[9] Mika V. Mantyla Daniel Graziotin, Miikka kuutila, The evolution of sentiment analysis, ITEE , University of Stuttgart, Finland.

[10] Dipankar Das, Souvick Ghosh and Tanmoy Chakraborty . Determining sentiment in citation text and analyzing its impact on the proposed ranking index. Jadavpur University, Kolkata. (references)