EasyChair Preprint

№ 2389

Survey Paper on Sentiment Analysis: Techniques and Challenges

Ansari Fatima Anees, Arsalaan Shaikh, Arbaz Shaikh and Sufiyan Shaikh

January 15, 2020

Survey Paper on Sentiment Analysis: Techniques and Challenges

Ansari Fatima Anees
Assistant Professor
Dept. of C.E.
M.H.S.S.C.O.E.
Mumbai, India
imfatima@gmail.com

Arsalaan Shaikh
Dept. of C.E.
M.H.S.S.C.O.E.
Mumbai, India
arsalaanshaikh1998@gmail.com

Arbaz Shaikh
Dept. of C.E.
M.H.S.S.C.O.E.
Mumbai, India
shaikharbaz655@gmail.com

Sufiyan Shaikh
Dept. of C.E.
M.H.S.S.C.O.E.
Mumbai, India
sufishaikh871998@gmail.com

***Abstract*—Process of finding out extracting experiences and emotions from the given dataset is called Sentiment Analysis. It is also called as Opinion Mining. By using sentiment analysis on the reviews the customer and enterprises can big a major change in the decision making process. There are different methodologies while making a sentiment analyzer. Data acquisition, data preprocessing and training with an algorithm are some of the steps involved in the methodology. There are various challenges while making a sentiment analyzer. In this paper we are going to survey different steps and techniques on sentiment analysis. We also studied previous work and tried to compare them and find out a better way to increase the accuracy and efficiency of a model. Naive Bayes and Support Vector Machine are mostly used classifiers. Further we discuss various challenges in sentiment analysis.**

***Keywords*—Sentiment Analysis, Opinion Mining, Lexicon based approach, Support Vector Machine (SVM), Naïve Bayes (NB), Product Reviews, Machine Learning.**

I. Introduction

With the rapid digitization over the last decade, people are using internet for performing all the basic tasks that they use to do manually traditionally. As e-commerce websites showcase the products with the help of images, it becomes difficult to know or understand the quality of product. Hence, because of this reason there is a feature of writing review and rating the products. Rating of products just displays the overall likeliness of product by the customers. The other feature i.e. reviews helps the potential user to understand the product from inside out. Since there are millions of users who access various e-commerce websites, there are lots of reviews available for each product. It becomes an impossible task for potential customer to read all the reviews even if it is possible, it is very difficult for a human to analyze those reviews. People often take reviews from their friends or relatives who have bought the product before buying it. In today's time reviews and ratings of the products plays major role to generate opinion.

To handle these problems Sentiment Analysis is used. Emotions of a sentence can easily understand using sentiment analysis. Automation of tedious tasks such as analyzing reviews can be done using Natural Language Processing. It can be carried out for many purposes such as detecting spam emails, finding inappropriate tweets, detecting fake reviews, finding features of the product form the reviews, etc. Finding the suitable techniques to analyse the reviews is the main objective of this paper . Information can be divided as subjective information and objective information. Objective information describes facts while subjective information deals with emotions. Subjective information is used for sentiment analysis.

Sentiment analysis can be carried out by two approaches i.e. machine learning approach and lexicon based approach. Fig 1. discusses classification of sentiment analysis in more detail.

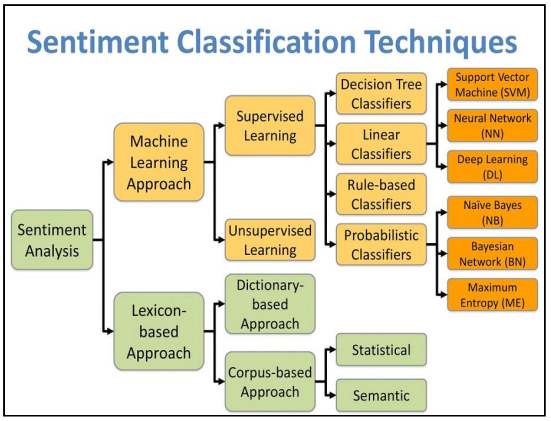

Fig. 1

II. METHODOLOGY

As mentioned previously, sentiment analysis uses two approaches i.e. machine learning approach and lexicon based analysis. Various steps are involved such as data acquisition, data preprocessing, lexicon based analysis, analyzed output.

A. Data Acquisition

Data acquisition is collection of the dataset for the analysis. The users will enter the product name available on an e-commerce website. We will use a web scraper to extract the reviews. Web scraper is a tool that is used to extract large amount of data from webpages and the process of extracting the data using web scraper tools is called web scraping. The web scraper used is beautiful soup. It is most suitable for small projects.

B. Data Preprocessing

The extracted reviews of a product are in the form of either sentences or paragraphs. The algorithm used here works on tokens or words hence data preprocessing is an essential step. Basically data preprocessing include tokenization, stop-words removal, stemming and lemmatization.

- Tokenization

It is a process of splitting or dividing the paragraphs into sentences and sentences into words. The tokenizer function that performs tokenization. There are two types of tokenizers: word tokenizer and sentence tokenizer. Word tokenizer tokenizes the sentence into words while sentence tokenizer tokenizes the paragraph into sentences as shown in Fig. 2 .

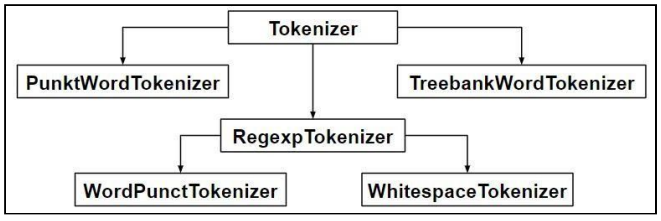

Fig. 2

- Stop-words Removal

Stop-words are words that are not considered or do not contribute in the analysis process. Rather than storing the stop-words in the datasets, it is better to remove the stop-words. Fig. 3 shows some examples on stopwords.

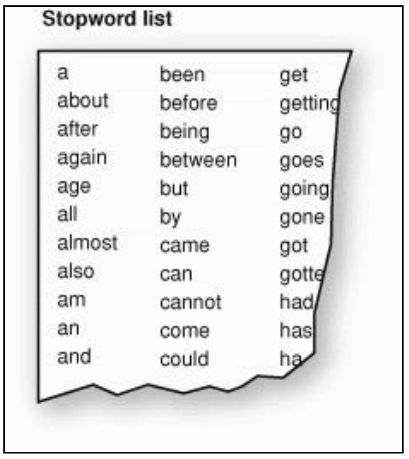

Fig. 3

- Stemming and Lemmatization

Stemming is the process of finding out morphemes or the root words from the derived word. It does not give the actual morphemes but the closest one.

Lemmatization is the process of finding out morphemes or the root words from the derived word. The lemmatization process gives the actual root word by comparing the stemmed words in its dictionary and gives the closest actual one. Fig. 4 describes the difference between lemmatization and stemming.

Word	*Lemmatization*	*Stemming*
was	be	wa
studies	study	studi
studying	study	study

Fig. 4

C. Analyzed Output

The input to data preprocessing step is the reviews. Output of the data preprocessing step is tokens. The token serve as an input to various algorithms which gives the output in the form of rating. These rating can be visualized in pi charts, bar graphs, etc.

III. SENTIMENT CLASSIFICATION TECHNIQUES

Lexicon based approach, Machine learning approach and Hybrid approach are the classification of Sentiment Analysis[1]. Lexicon based approach divides the entire document into lexemes which is used to examine the sentences. Lexicon based approach is further classified as corpus based approach and dictionary based approach. Corpus based approach finds out the polarity of the sentence as negative, positive and neutral. Positive – beautiful, best, excellent, etc. and Negative – bad, disgusting, irritating, etc. Dictionary based approach is a mathematical approach for

measuring the feeling that the sentences conveys to the reader.

Machine learning approach uses machine learning algorithm and it is classified as supervised learning and unsupervised learning. Supervised learning requires desired output to compare with the actual output. While unsupervised learning does not require any desire output rather it uses previous experience and data to improve its accuracy. Machine learning approach and Lexicon based approach combines to form Hybrid approach..

A. Lexicon Based Approach

As mentioned above, lexicon based approach basically deals with lexeme i.e. tokens or words. It splits the sentence into tokens and processes them. These words are classified as positive or negative opinions.

Classification of Lexicon based approach is as follows:

- Corpus based approach

It came into the picture to resolve the problems of dictionary based approach. It is less efficient than dictionary based approach because there is a need to make a large corpus for covering English words which is a difficult task.

- Dictionary based approach

It provides better efficiency than corpus based approach. It uses a dictionary which consists of all the synonyms and antonyms of each words. It cannot find the opinion with domain specific orientation.

B. Machine Learning Approach

It uses different types of algorithm to carry out the sentiment analysis. It includes training the particular portion of dataset and then using the remaining portion of dataset to test for the result. Majorly used algorithm is:

- Naïve Bayes

Naïve Bayes algorithm is derived from Bayes' theorem. It consists of a family of algorithms. Bayes' theorem computes the probability of given set using already calculated probabilities. Fig. 5 describes Bayes' Theorem mathematically,

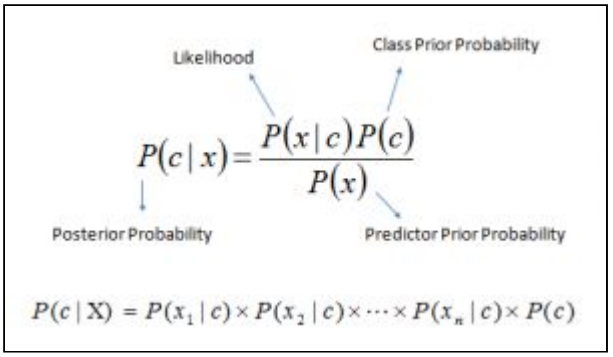

Fig. 5

IV. Related Work

Till now, many researchers have worked on the analysis of product reviews. In paper [3] Elli, Maria and Yi-Fan extracted sentiments from the reviews and analyzed the result to build a business model. They claimed that the proposed tools gives high enough accuracy due to its robustness. They made their decision more accurate by using business analytics. Other notable work includes emotion detection, gender based on the names and fake review detection. The commonly used programming languages are Python and R. Monomial Naïve Bayes along with Support Vector Machine were used. In paper [4] supervised learning algorithms are used to generate ratings on a numerical scale with the help of text only. Their total dataset is divided into 70% training data and 30% testing data. They used various classifiers for their module.

The author in Paper [5] continues the idea of Sentiment Analysis and Natural Language Processing. Reviews are labeled as positive or negative using Naïve Bayes' and decision list classifier. The author in paper [6] builds a system that helps to visualise the sentiment in the form of charts. Scraping was used to collect usable data from product's URL. Author generated sentiments using Naive Bayes and Support Vector Machine algorithms. The paper does not mention accuracy anywhere and the output is represented in statistical form. In the paper [7] writer predicts product rating based on concept of bag-of words. These models utilizes unigrams and bigrams. The Amazon video game reviews subset of UCSD Time Based model didn't work well, since the variance in the average rating between each day, month or year was relatively small. Unigram produce better results as compared to Bigram. Unigram results had an approximately 16.00% better performance than bigrams.

In paper [8] various feature extraction or selection techniques for sentiment analysis are performed for the analysis purpose. The collected dataset from Amazon is preprocessed before passing through classifier. They carried out tokenization, stemming, lemmatization and stopwords removal techniques in data preprocessing step. They only used Naïve Bayes classifier which did not produce sufficient output or results. Paper [9] uses an easier algorithm for understanding the sentiments. Techniques such as Support vector machine (SVM), logistic regression and decision trees were used in this paper. The SVM algorithm has a high accuracy for small dataset but it does not perform well on relatively large datasets. In paper [10] TF IDF was used for experiments on dataset. Rating of reviews are predicted with the help of bag of words. Classifiers like root mean square error and linear regression model were used. Above mentioned are some related works on sentiment analysis and techniques. We will try to build a more accurate model by considering the above mentioned works.

V. Challenges

A. Multiple Language Input

As the dataset is a collection of the reviews given by the users, it can be in multiple languages. But the classifier mainly uses English language. Therefore it becomes very difficult to train the algorithm for different languages other

than English. Hence Multiple Language Input is a big challenge in sentiment analysis.

B. Fake Inputs

Fake or bogus reviews misguide the users or customers about a product by providing fake or irrelevant negative or positive reviews. This is mostly done to increase or decrease the popularity of a product. So, identifying fake reviews is a tedious and almost impossible task.

C. Emoticons and Sarcastic Reviews

Emoticons are the pictorial representation of one's expressions. Using emoticons to define the product makes it easier for the customer or user to understand one's feelings. On the other hand it becomes difficult for a machine to understand the emoticons. It is not easy to train an algorithm for emoticons as an input.

Sarcastic reviews are difficult to interpret by the machine. The model needs to be trained with more and more such data to give an accurate answer. Hence, Emoticons and sarcastic reviews are one of the biggest challenges of sentiment analysis.

VI. Conclusion

Sentiment analysis is a growing domain in today's digital world. In this paper we have collected data in the form of reviews from amazon e-commerce website. Web scraper was used to scrape reviews from amazon product url and saved in the form of spreadsheet. The scraped reviews are preprocessed to save computational time and storage. Since sentiment analysis involves three approaches anyone can be used since each one has its pros and cons. We have used lexicon based approach to compute the sentiment of the reviews. Lexicon based approach gives better accuracy as they make use of a dictionary. As mentioned above it is very difficult to deal with the challenges of sentiment analysis.

VII. References

[1] Diana Maynard, Adam Funk. Automatic detection of political opinions in tweets. In: Proceedings of the 8th international conference on the semantic web, ESWC'11; 2011.

[2] (2015) Machine Learning with Naïve Bayes Classifier [Online].Available:http://blog.datumbox.com/machine-learningtutorial-the naive-bayes-text-classifier/

[3] Elli, Maria Soledad, and Yi-Fan Wang. "Amazon Reviews, business analytics with sentiment analysis." 2016

[4] Xu, Yun, Xinhui Wu, and Qinxia Wang. "Sentiment Analysis of Yelp„s Ratings Based on Text Reviews." (2015).

[5] Rain, Callen. "Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning."Swarthmore College (2013).

[6] Bhatt, Aashutosh, et al. "Amazon Review Classification and Sentiment Analysis." International Journal of Computer Science and Information Technologies 6.6 (2015): 5107-5110.

[7] Chen, Weikang, Chihhung Lin, and Yi-Shu Tai."Text-Based Rating Predictions on Amazon Health & Personal Care Product Review." (2015).

[8] Shaikh, Tahura, and DeepaDeshpande. "Feature Selection Methods in Sentiment Analysis and Sentiment Classification of Amazon Product Reviews.",(2016)

[9] Nasr, Mona Mohamed, Essam Mohamed Shaaban, and Ahmed Mostafa Hafez. "Building Sentiment analysis Model using Graphlab." IJSER, 2017

[10] Text mining for yelp dataset challenge; Mingshan Wang; University of California San Diego, (2017)

[11] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Spring