International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET21834

Social Media Behavioral Intelligence using Feature Extraction

Panchal Mayuriben*, Dr. Priyanka Sharma
School of Information Technology, Artificial Intelligence, and Cyber Security, Rashtriya Raksha University, Gandhinagar, Gujarat, India

Article Info
Volume 8, Issue 3
Page Number: 01-06

Publication Issue :
May-June-2021

Article History
Accepted : 01 May 2021
Published: 08 May 2021

ABSTRACT

Analysis of the behavioral pattern of a people using data of the social media became a trend in last couple of years. Among this popular network, Twitter, Facebook and the Instagram become more and more popular and that's why these platforms attract the lots of researchers to predict the sentiment regarding major events like election, product brand, movie, stock market and recent trends are some of them. By identifying the attitude associated with the text in terms of positive, negative or the neutral we are able to analyze the opinion behind the content generated by the user and this opinions about the sentiment are very helpful to for the organization or the political parties or among other entities. The task of sentiment analysis is conducted using identifying the polarity associated with the word or document or we can say sentence. This paper consists research work which is designed to improve the accuracy of the model by improving the Naïve Bayes algorithm and I also worked to improve the 3-gram method during my research

Keywords: Web Content Mining, Sentiment analysis, Opinion analysis, Clustering, Classification.

I. INTRODUCTION

Sentiment Analysis is a technique used in the field of a text mining which finds the probabilities of different classes which are assigned to a particular text. The outcome of sentiment analysis is in terms of the joint probabilities of particular word and the classes associated to that words. Social Media Sentiment Analysis is nothing but a set of advance mining techniques which are able to analyse the sentiment associated with the given text or word. The primary goal of these sentiment analysis algorithms is to identify the sentiment in the form of positive, negative or neutral one. The process of sentiment analysis is sometimes known as an Opinion Mining which have a primary goal of analysing the given conversations or opinions of the tweets or comments or reviews. This analysis is very useful to make business decisions or make a particular strategy for the organization. There are some well-known tools like Enginuity, Steamcrab, MeaningCloud are available in the market for the analysis purpose of the available data of a social media. The languages like R and Python are used for the sentiment analysis of the social media dataset.

II. OBJECTIVES

The proposed system consists of 4 modules namely Retrieve tweets or comments, Pre-Processing, Sentiment Score and Review of sentiment.

A. Data Retrieval

There are mainly three methods are there to collect the data which consist Repository of the data, Automated tools and by using the Premium tools. Generally, data repository like UCI and the SNAP are used to collect the tweets of the user. We can also use APIs such as Stream API and the Search API to collect the data directly from the system. Use of Stream API is to collect data directly from the twitter while on the other hand search API is used to collect data based on the hash tag used in the tweet.

B. Pre-Processing of Data

As we know that data collected via any tool is a raw data and we have to process that data using various pre-processing techniques and it involves the task such as eliminating the stop words as well as notation used in the data. We also have to remove the URLs as well as emoticons used in the tweets. After the pre-processing stage we can feed that data to the classifier for further classification of that data.

C. Defining A Sentiment Score

Sentiment score is assigned by the AFINN dictionary. Two versions of AFINN dictionary are AFINN-96 and AFINN-111. AFINN-96 consists 1468 words and AFINN-111 consists 2477 words including 15 phrases.

AFINN dictionary is responsible for assigning the sentiment score to the given word. There are mainly two AFINN dictionary is available and apart from that AFINN-96 includes 1468 words while AFINN-111 includes 2477 words. It generates basically four types of sentiment based on the output score. If the score ranges between negative 5 to negative 4 then it assigns sentiment called “very negative”, if the score is in between negative 3 to negative 1 then it assign sentiment called “Negative”, same if the score is in between positive one to positive 3 then it assign sentiment as a “Positive” and if the score ranges from positive 4 to positive 5 then it assign sentiment as “Very Positive”

D. Sentiment Review

By the combination of supervised and unsupervised learning methods we are able to extract the review of the sentiment in terms of polarity of the word for a given sentence. Supervised learning focuses on quantitative data and uses classifier like SVM and NB to extract the data while on the other hand unsupervised learning focuses on qualitative data and uses clustering technique like FCM to extract the data.

III. PROBLEM STATEMENT

Here by studying various research work i found some major problems in the current methodology of the recommendation and sentiment analysis which I listed below.

Various data preprocessing methods and during the brief study of my research I found various disadvantages of these methods which I listed below.

- Aggregation method generally combine two or more attribute into a single one but fail to co-relate the different attribute so, it gives a devastating result in a relational database. [1]
- Simple Random sampling method assumes equal probability for all the items which leads to a negative impact on a survey
- Dimensionality reduction reduce the dimensionality of a data set but it also fails to combine co-related data

The Probability based Feature Extraction method select prominent features from the high dimensional

feature vector by eliminating irrelevant features which improves performance of the system [2].

During the study of previous research paper, I found that clustering using SVM and Fuzzy C-Means doesn't perform well due to particular reasons which are mentioned below

- SVM doesn't perform well when we have large data set because the required training time is higher [3].
- Fuzzy C-Means doesn't perform very well, when the data set has more noise i.e. target classes will overlap [4].

IV. BACKGROUND THEORY

Extraction process of the opinion given by the public on a social media is a challenging task and additionally social media has a rich content which needs to explore using the computation model of the NLP which again motivates researchers to improve the existing algorithms. In last decade, numbers of research are conducted on the sentiment analysis based on the reviews, opinion, comments and the articles about how sentiment are expressed in an informal manner. Previous study suggest that Twitter data has an excellent predictability power in the are of stock market to the election conducted in the different countries.

E Junqué de Fortuny, T De Smedt, D Martens 2014 proposed a method to collect the identical text from the internet and on that text, they perform the sentiment analysis. The main study of area is the Belgian elections which was conducted in the year 2010. They build the web crawler for analysing the data of the election. Mining module was developed using python language. Module consist of more than 3000 lexicons of a dutch adjectives which are occurred in the given article. Then they assign a polarity score to each adjective manually. Biases as well as sentiment associated with the word is analysed through the entire document. The sentiment score associated with every party was calculated on a weekly basis and then they took a simple moving average to smoothen any fluctuations associated with the individual party and then model calculates the difference associated with the different parties.

L Chen, W Wang, M Nagarajan, S Wang, AP Sheth - ICWSM, 2012 present a paper which is based on an optimization technique that extract the information regarding sentiment for a given text. There is a diverse range of expressions for the given sentiment are recognized using this technique. Target dependent polarity associated with the tweet sentiment is identified. A set of words from the traditional as well as language called slang are obtained with help of lexicon dictionary like SentiWordNet, MPQA as well as Urban Dictionary. Here they take assumption that target is already identified in the given tweet. Here sentiment word is extracted from the identified target. After this process is completed relation of consistency and inconsistency will be identified and established for the further use to build the network. In this model Edge represents the consistent and inconsistent relation between the given node and weight represents the frequency of the given word regarding the sentiment. Polarity probability is calculated using an optimization technique. Here Positive and Negative sentiment is calculated based on the consistent and inconsistent relationship.

Johan Bollen, Huina Mao, Alberto Pepe (2011), define a solution for the mood detection using a sentiment analysis in which they use correlation between the political culture, Socioeconomic events as well as mood of the public using their tweets. The data set includes approximately 9.6 million tweets published in the corpus between August to December 2008. They used psychometric instrument named as Profile of Mood State which includes mood like tension, anger, vigor, confusion, depression to extract the information. Then after tweets are scored with POMS

scoring function to count the number of predefined adjectives for each mood dimensions. Z-score method is used to normalized the generated disparity of collected tweets.

V. EXISTING WORK

Sentiment Analysis is perform through NLP (Natural Language Processing) which strats from the document level classification (Turney, 2002; Pang and Lee, 2004), to the sentence level (Hu and Liu, 2004; Kim and Hovy, 2004) as well as recently it is perform at a phrase level (Wilson et al., 2005; Agarwal et al., 2009).

Data of a Microblogging sites like Twitter able users to post various reactions and opinions about the current scenarios as well as about the different challenges faced by the people. Some of the early sentiment analysis techniques uses a distant learning methodology to find the sentiment data in which they use tweets which ends with either positive emoticons as well as negative emoticons Pak and Paroubek (2010). Authors report that techniques like POS and the Bigrams are useful and uses n-gram models. Moreover, the data they use for training and testing is collected by search queries and is therefore biased.

More significant approach for the classification of the sentiment is conducted by the Barbosa and Feng (2010).They apply polarity predictions from three different websites and use more than 1000 manual labled tweets for the tuning purpose and also use another 1000 manual labled data for the testing purpose. They proposed the feature called syntax feature of tweets and hash tag links as well as punctuations and POS words.

Gamon (2004) applies a sentiment analysis on the feedback data of the global support service survey. The main aim of this paper is to analyze the role of the POS tags by performing the feature analysis and the feature selection techniques and then to demonstrate the linguistic analysis on that feature to improve the classification accuracy.

Sentiment analysis has been handled as a Natural Language Processing task at many levels of granularity. Starting from being a document level classification task (Turney, 2002; Pang and Lee, 2004), it has been handled at the sentence level (Hu and Liu, 2004; Kim and Hovy, 2004) and more recently at the phrase level (Wilson et al., 2005; Agarwal et al., 2009).

VI. PROPOSED SYSTEM

System flow of the proposed system is given in below diagram with all the modules and functions performed by the system.

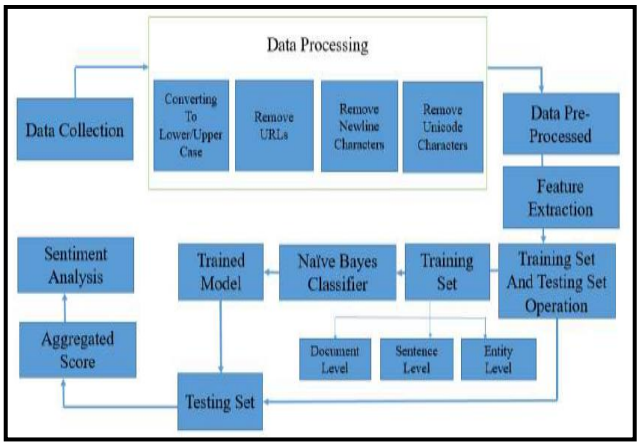

Fig 1. Proposed System Flow

As we know we divide our project into three different parts we will conduct our further study in four different parts which are **data pre-processing, feature extraction, clustering** and **classification**

A. Data Pre-Processing

As we know that this process consists of large data so, first of all we have to collect data from a large data set and from this data set we have to again clean the data which is required for our project by applying Knowledge Discovery Process which is iterative and

interactive method and consist nine major sub processes which I listed below.

i) Data understanding with application domain

ii) Selecting and creating dataset

iii) Preprocessing and cleaning

iv) Data Transformation

v) Data Mining Task selection

vi) Choosing the Data Mining algorithm

vii) Employing the Data Mining algorithm

viii) Pattern evaluation

ix) Using the discovered knowledge.

B. Extraction of Feature

Here I used mainly four extraction methods to extract the features from the sentence which are given below.

Probability Based Feature Extraction: In this feature we calculate basic Weighted probability of a given word for a particular category and it is given by the following formula.

(weight*assumed probability + total number of appearances * basic probability) divides by (total number of appearances + weight)

Default weight is taken as a 1 and the assumed probability is chosen as a 0.5 to calculate the weighted probability.

Emoticons: This feature predict the appropriate emoticons as per the text entered like it provides different kind of emoticons for the happy, sad or any other mood. It uses lexicon directory to predict appropriate emoticons and here I combine Lexicon directory with the AFINN directory to get more accurate result.

Synonyms: If there is a word which is not available in the directory, I used then it will provide zero as the sentiment score which degrade the accuracy of the result and to resolve this problem, I used synonyms in the system which replace the word which is not available in the directory with the word in a directory which have a same meaning as the given word.

n-gram: N-gram provides a sequence for the items used in the model as a form of speech or in the form of text. There are many techniques like unigram, bigram or trigram used to arrange the sequence of data in the model. I used trigram technique to improve the performance of the model which identifies the difference between the sentence which are positively said to the sentence which are negatively said.

Here we also Utilize Bag of Words Representation with custom Bigrams and Trigrams [5].

C. Classification

During the study of previous research paper, I found that clustering using SVM and Fuzzy C-Means doesn't perform well due to particular reasons which are mentioned below

- SVM doesn't perform well when we have large data set because the required training time is higher.
- Fuzzy C-Means doesn't perform very well, when the data set has more noise i.e. target classes will overlap.

So, to overcome this problem I, design a new solution by using Naive Bayes Classifier via **nltk** Package [6] in Python which Predict Sentiment (-1,0,1) of each tweet in a text file by using following three functionalities

- Removing feature List duplicates
- start get feature Vector using stop words
- Creating Features vector list

By using these three features we get F1 score

After that we use Naïve Bayes algorithm to segregate one class from the other class which is given below

Naive Bayes implementation:

$p(A|B) = p(B|A) * p(A) / p(B)$

Assume A - Category

B - Test data

p(A|B) - Category given the Test data

Here ignoring p(B) in the denominator (Since it remains same for every category)

VII. CONCLUSION

The improvement in data collection and use of that data to figure out useful insights is a very crucial task to get the mood of the people and in this project, I try to utilize all the parameters for the finding the useful information from the available data. By completing my research study, I reach on a conclusion that sentiment analysis is a very useful technique for defining user behavior which is useful for taking important decision at a business level as well as social level. Decision making will be very easy by studying historical data of the particular user or a user group.

VIII. REFERENCES

[1]. M. H. Uma K, "Data Collection Methods and Data Preprocessing Techniques for Healthcare Data Using Data Mining," International Journal of Scientific & Engineering Research, vol. 8, no. 6, p. 2, 2017.

[2]. N. M. Basant Agarwal, "Categorical Probability Proportion Difference (CPPD): A Feature Selection Method for Sentiment Classification," 2012.

[3]. A. J. Reagan, "Sentiment analysis methods for understanding large-scale texts: a case for using continuum-scored words and word shift graphs," 2017.

[4]. A. D. A. J. a. C. T. R. Mayank Gupta, "Sentiment Analysis in Twitter".

[5]. A. S. Ankush Mittal, "SENTIMENT ANALYSIS USING N-GRAM ALGO AND SVM CLASSIFIER," vol. 5, no. 4, 2017.

[6]. A. M. D. A. Mohamad Syahrul Mubarok, "Aspect-based sentiment analysis to review products using Naïve Bayes," 2017.

Cite this article as :

Panchal Mayuriben, Dr. Priyanka Sharma, "Social Media Behavioral Intelligence using Feature Extraction", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 8 Issue 3, pp. 01-06, May-June 2021. Available at

doi : https://doi.org/10.32628/IJSRSET21834

Journal URL : https://ijsrset.com/IJSRSET21834