Behavioural Analysis of Tweeter data : A Classification Approach

Chithra R G
Department of Computer Science and Engineering
Alva's Institute of Engineering and Technology
Moodbidri, India

Harshitha G M
Assitant Professor
Department of Computer Science and Engineering
Alva's Institute of Engineering and Technology
Moodbidri, India

AnuPrakash M P
Department of Computer Science and Engineering
Alva's Institute of Engineering and Technology
Moodbidri, India

Rakshitha H B
Department of Computer Science and Engineering
Alva's Institute of Engineering and Technology
Moodbidri, India

***Abstract*—— The tremendous interest in the social network sites among Internet users is growing exponentially with respect to the raising in Technology rate in the World. Twitter is the social networking service on which user's posts and interact with the messages called as "tweets". Every year millions of tweets are posted in this site and this leads to analyze the mood of the user based upon his/her respective post. In this analysis the system will the behavior of the person like passive, aggressive and passive-aggressive by considering the total number of positive, negative or neutral opinion of the user in all his/her posts. The abundance of social media data provides opportunities to understand criminal minded experiences, but also raise methodology difficulties in making sense of social media data for educational purposes. Behavior analysis are made on the stored tweet repository by considering several algorithms and the classification are made between the tweets. The main objective is to avoid the viral or unwanted or scary tweets from the media and this will helps to give the best recommendation towards the positive users through their experience in the social site.** **Keywords—tweets, types of opinion, behavior analysis, repository, classification method.**

I. INTRODUCTION

Social media has become an essential part of life, it is a comparatively cheap and widely accessible medium that enables anyone to broadcast and access information / news / knowledge, and to build new relations. It is a tool which is used to share different opinions that can belong to different subjects; such as humanitarian causes, environmental irregularities, economic issues, or political disputes. Some of the social media allows the users to interact with other who are far away from them. Due to simple and easy privacy policies and easy accessibility of the social media the peoples are wills to use more and more. Twitter is one of the social media people used to share their opinion through their post. Twitter is a social network which allows its users to post and share short messages (up to 140 characters) called tweets. Over the past decades, Twitter has spread worldwide and has become one of the major social networks. Behavioral analysis is the type of the analysis where the data from the set of data stored in the data repository are taken and identifying the behavior or the opinion of the user based on that data. Behavioral analysis is the type of the analysis where the data from the set of data stored in the data repository are taken and identifying the behavior or the opinion of the user based on that data. In this type of analysis many algorithms under the data mining are used. Behavioral analysis mainly considers the attributes such as: mood, thinking, emotion, communication, and socialization. This analysis is used to design the classifiers to identify the passive, aggressive and passive-aggressive behavior of the users. In this type of analysis many algorithms under the data mining are used.

II. RELATED WORKS

Behavioral analysis mainly considers the attributes such as: mood, thinking, emotion, communication, and socialization. This analysis is used to design the classifiers to identify the positive, negative and neutral users.

In paper [1] the authors, Varsha Sahanayak et al., introduce the automatic method of classifying the sentiments of Tweets taken from Twitter dataset. The authors explain the analysis of the data using the machine learning approach. Different methods are used in this approach like Naive Bayes, Maximum Entropy (Maxent), and Support Vector Machines (SVM) are the machine learning classifiers. The proposed is based on two important parts viz Data Extraction, preprocessing of extracted data and classification. To uncover the sentiments, the system will first extract the opinion words from tweets and then we find out their orientation, i.e., to decide whether each sentiment word reflects exaggerated and self-indulgent feelings of tenderness, sadness, or nostalgia. The different steps are used in this approach.

In paper [2] the authors explain the key characteristics of the tweets like maximum length of the tweeter is 140 character, writing technique, amount of data available, topics that should be mentioned in the data and the real time. Basic characteristics of the twitter is explained like hashtag, emoticons, and target and special symbols. The paper also explains about the methodology with different steps used to analyze the data. Steps like pre-processing of tweets, scoring models and tweet sentiment scoring.

In paper [3] the authors, Ms. Umaa Ramakrishnan et al., explains about the methodologies which can be used in machine learning method to analyze the data like lexicon-based approach and Statistical based approach. The paper also explains about the classifiers which are used for the analysis of the data like Naïve-Bayes Classifier, Bayesian Network, Maximum Entropy Classifier, Linear Classifiers, Decision Tree Classifiers and Rule-based Classifiers. The authors mainly explain the use of the natural language processing in analyzing the data. The main idea behind NLP is that we must analyze the data set by accepting the data set and reading the data.

In paper [4] the paper mainly explains about the need of the analyzing social media data using the different techniques like sentiment analysis, text mining, machine learning, KNN algorithm and feature selection mode. The authors explain about the method of analyzing the data with different steps like data is extracted from the tweets or the set of data sets which is stored online in the form of repository. Pre-processing of the data using the different steps 1. Tokenization 2. Stop word removal 3. Stemming 4. Weighting Factor 5. Term-document matrix 6. Term-Document matrix 7. Creating dictionary of terms. KNN algorithm with explanations and its applications are explained. Naïve Bayes Algorithm is also explained in detail.

In paper [5] the authors explain about the social networking and micro blogging service that allows the user to post real time messages called tweets. In paper brief terminology associated with tweets are mentioned they are emoticons, target and hash tag. In pre-processing of tweeter data two new resources are introduced they are 1. An emoticon dictionary and 2. An acronym dictionary. The paper uses the dictionary called Dictionary of Affect in Language (DAL) and extend it using Word Net. Design of Tree Kernel is introduced where it is used to combine many categories of feature in one form of representation.

In paper [6] the authors, Pierre Ficamos et al., explains about the feature extraction methods that are mostly specified in account of data analysis. They are 1. Presence is better than Frequency 2. Negation Handling 3. Bigrams 4. Part of Speech (POS) tags and 5. Steaming word. Data Processing technique explained with explain the set of removed bag of words Lower uppercase letters, remove digits, remove stop words, remove repeated letters, Tokenize, Detect POS tags, Lemmatize. At last methodology is explained with the help of topic extraction and training the algorithm to remove the extra bag of words from the dataset.

In paper [7] the authors explain the system architecture and the different methods used for analyzing the data present in the data set. The methodology present in the paper consist of Getting Data from Twitter Streaming API and Predictive Modelling. Where in Getting Data from Twitter Streaming API- API makes the interaction with computer programs and web services easy. Web services provide APIs to interact with their services. To establish Twitter Streaming API the system, need 4 pieces of information: API key, API secret, Access token and Access token secret and Predictive Modelling is made up of a number of predictors, which are variable factors that are likely to influence future behavior or results. The predictive modelling approach uses linear regression with the following parameters: customer's gender, age, purchase history, and future sale.

III. PROPOSED SYSTEM

The proposed system uses data extracted from social networking web sites, like tweeter. The data can be exacted by the tweeter in the form of excel sheet. The main objective of the project is to analyze the tweeter data and to find the behavior of the user. The proposed system uses the different modules like tweeter import process, tweeter preprocessing, self-learning system and sentimental analysis. Based on the number of positive, negative and neutral tweets in the input the behavior of the person can be classified as passive, aggressive or passive-aggressive.

A. *System Design*

The design of the system architecture describes the structure, behavior and more views of the system and analysis. The goal of design is to produce a module of the system which is used to build the system. In the proposed system.

The input for the process can be given in the form of excel file or in the text format. The input which is given in the form of text format is added to the excel file. The data present in the excel file is imported and is stored in the database. The will input the specific keyword on which the analysis should be made. Keyword is compared with data present in the database and the related data is retrieved for the analysis. This process is done using the key word identification algorithm iteratively.

The retrieved data is undertaken to the pre-processing step where URL’s, special characters, unnecessary words are removed from the data and stored in the database in separate table. Stemming process is also made where the prefix and the suffix of the words are removed and only the main is stored in the database. The use of the stemming process is that it reduces the storage space. This two process is done using the pre-processing algorithm.

If the unknown word or the word which is not present in the database occurs repeatedly in the input it can be added to the database dictionary directly. Threshold value is set in the process to add the word directly. In the propose system threshold value is set to 5. If in the input the specific is repeated more than or equal to threshold value it will be directly added to the database. This process is done using the auto-inclusion of the sensitive words or the self-learning algorithm.

The pre-processed data are sent to the analysis process where initially the value of positive, negative and the neutral are set to zero. The words are compared iteratively with the dictionary present in the database. Sentiment analysis can be considered as the classification process, where it classify the data into positive , negative or neutral on which the behavior of the user can be analyzed. Based on the number of positive, negative or the neutral tweets in the input the behavior of the person can be classified as passive, aggressive or passive-aggressive.

B. *System Architecture*

The goal of design is to produce a module of the system which is used to build the system. Fig 1 shows the proposed system where:

- The tweets can be uploaded in the form of excel file, user input or directly by the tweeter for the specified topic.
- Tweets present in the excel file are imported to the database and is stored in the database.

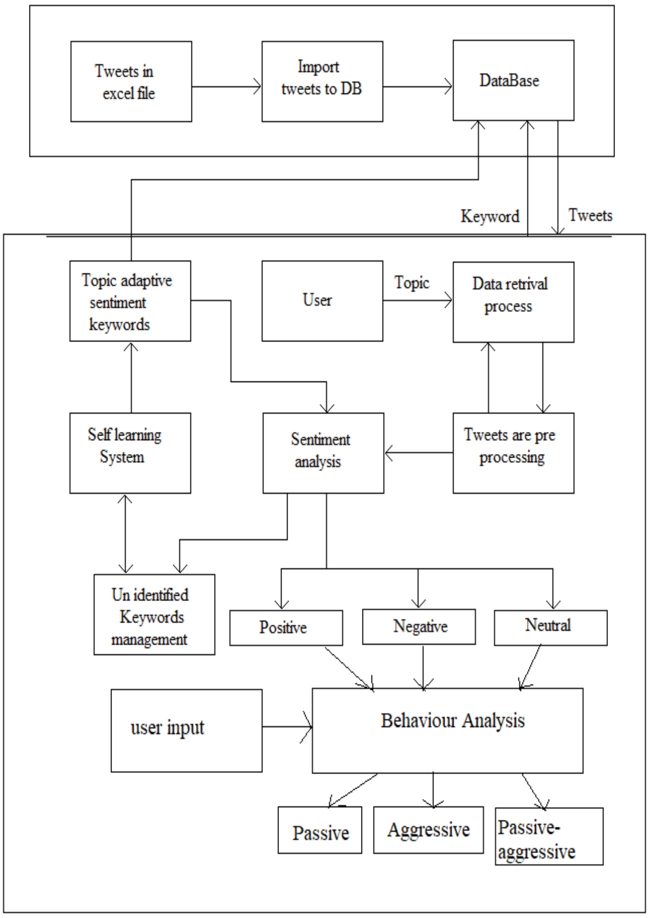


Fig 1: System Architecture.

- The user will input the topic on which the analysis should be made and that topic is called as keyword.
- Keyword is sent to the database to retrieve the related tweets present in the database.
- After retrieving the tweets, they are pre-processed using the pre-processing algorithm.
- The tweets given as the input by the user is stored in the database and they are also retrieved.
- Threshold has been set to add the word that are not present previously in the database.
- During the input by the user if any of the specific word is not in the database they are added to the database this is done using the self-learning system.
- The pre-processed keywords are analyzed using the sentiment classification algorithm.
- Then the tweets with the sentiment are classified as the positive, negative and neutral tweets.
- The left out word which is not in any sentiment type, that word sentiment is decided based on the positive ,negative and neutral count in that particular tweet. If positive count is more than negative and neutral them it will be considered positive sentiment only.
- Based on the sentimental analysis of the input, person's behavior is classified into passive, aggressive or the passive-aggressive.
- In the input if the positive count is high than it is considered as the passive behavior, if the negative count is high it is considered as the aggressive and if the positive count and the negative counts are equal than it is considered as the passive-aggressive.

C. *Algorithm*

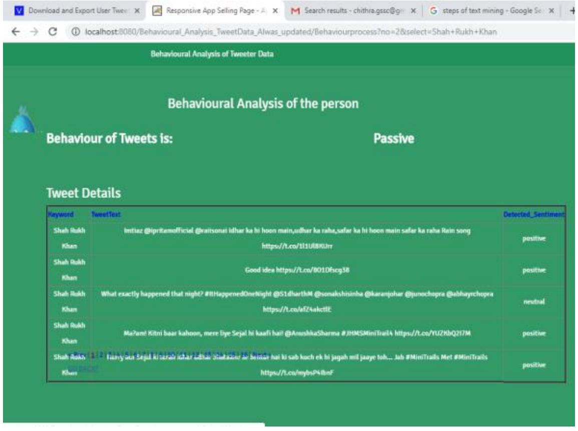
Algorithm can perform calculation, data processing, and automated reasoning tasks. As an effective method, an algorithm can be expressed within a finite amount of space and time and in a well-defined formal language for calculating a function. Sentiment Classification algorithm is used to classify the data into positive, negative and neutral tweets. Then the tweets are imported to the database from where the data for future process is retrieved.

```
Algorithm 1: Sentiment Analysis Algorithm
Input: The pre-processed data present in data base.
Output: The analyzed tweet with positive, negative, neutral outputs.
Step 1: Let n be number of tweets
Step 2: for i=0 to n
        Count positive=1, negative=1, neutral=1.
                Count= Count+1
Step 3: Based on positive, negative, neutral count
Step 4: Give Sentimental process
Step 5: end for loop
```

The number of tweets present in the excel file can be considered as 'n'. Initially positive, negative and neutral are initialized as '0'. The algorithm will check the data iteratively for n times. If it founds the positive, negative or the neutral tweet its count will be incremented by '1'. Based on the count of the positive, negative and neutral tweets the result will be displayed. After the pre-processing the data will be sent to the sentiment analysis algorithm to analyses the behavior of the person. Using the keywords selected by the user the analysis will be made for that data and the number of tweets, positive tweets, negative tweets, the neutral tweets and also number of unmanaged tweets are shown as the result of the sentimental analysis. The overall system will show the behavior of the person like passive, aggressive or the passive-aggressive based on output of the sentimental analysis. And the result will be shown as passive, aggressive or passive-aggressive.

IV. RESULTS

The application is used to classify the behavior of the person as passive, aggressive or passive-aggressive. In the present system only the sentimental analysis is done where the input data is classified as positive, negative and natural. The proposed system helps us to analyze the tweets present in the input and to find the behavior of the person. The below figure shows the example where the input data has been classified as passive by considering the sentimental analysis input.


Download and Export User Twee
Responsive App Selling Page - A
Search results - chithra.gssc@g
steps of text mining - Google Se

localhost:8080/Behavioural_Analysis_TweetData_Alwas_updated/Behaviourprocess?no=2&select=Shah+Rukh+Khan

Behavioural Analysis of Tweeter Data

Behavioural Analysis of the person

Behaviour of Tweets is: Passive

Tweet Details

Keyword	TweetText	Detected_Sentiment
Shah Rukh Khan	Imtiaz @ipritamofficial @raitsonai idhar ka hi hoon main,udhar ka raha,safar ka hi hoon main safar ka raha Rain song https://t.co/1l1Ul8KUrr	positive
Shah Rukh Khan	Good idea https://t.co/8O1Dfscg38	positive
Shah Rukh Khan	What exactly happened that night? #ItHappenedOneNight @S1dharthM @sonakshisinha @karanjohar @junochopra @abhayrchopra https://t.co/afZ4akctlE	neutral
Shah Rukh Khan	Ma?am! Kitni baar kahoon, mere liye Sejal hi kaafi hai! @AnushkaSharma #JHMSMiniTrail4 https://t.co/YUZKbQ2l7M	positive
Shah Rukh Khan	Harry aur Sejal ki tarah idhar udhar [illegible] hai ki sab koch ek hi jagah mil jaaye toh... Jab #MiniTrails Met #MiniTrails https://t.co/mybsP4lbnF	positive

Fig 2: Experimental result

The figure shows the experimental result where it shows the keyword on which the analysis is done, the tweet text and the specified sentimental analysis result of that tweet. By considering the output of the sentimental analysis the behavior of the person is classified as passive, as there is more number of positive tweets in the input file.

V. CONCLUSION

It has been found that many methods are used to analyze the behavior of tweets, but the proposed system uses the iterative method where the storage space can be reduced and the analysis of the data can be made easily. The data can be classified accurately into the positive, negative and neutral data and the system can also achieve the efficiency and based on that behavior of the person can be classified.

REFERENCES

[1] Sentiment Analysis on Twitter Data: Varsha Sahayak, Vijaya Shete, Apashabi Pathan. International Journal of Innovative Research in Advanced Engineering (IJIRAE), Volume2, January 2015.

[2] Sentiment Analysis on Twitter: Akshi Kumar and Teeja Mary Sebastian. (IJCSI) International Journal of Computer Science Issues, Vol. 9, No 3, July 2012.

[3] Sentiment Analysis of Twitter Data: Based on User Behavior: Ms. Umaa Ramakrishnan and Ms. Rashmi Shankar. International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 7 (2015) pp. 16291-16301.

[4] Approaches for Sentiment Analysis on Twitter: A Stateof-Art study Harsh Thakkar and Dhiren Patel.

[5] A Topic based Approach for Sentiment Analysis on Twitter Data: Pierre FICAMOS, Yan LIU. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 12, 2016R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Sentiment analysis in twitter data using data analytic techniques for predictive modelling: A Razia Sulthana, A K Jaithunbi, L Sai Ramesh. (2018)

[7] Sentiment analysis in twitter data using data analytic techniques for predictive modelling: A Razia Sulthana, A K Jaithunbi, L Sai Rame