SENTIMENT ANALYSIS OF SOCIAL MEDIA DATA USING DEEP LEARNING

G MAHESH CHALLARI[1], K. VIJAY KUMAR[2]

[1] Assistant Professor in Department of cse , R & D Co-ordinator at Sree Dattha Institute of Engineering And Science, Sheriguda, Hyderabad, India, 501510
[2]Assistant Professor in Department of cse at Sree Dattha Institute of Engineering And Science, Sheriguda, Hyderabad, India, 501510

---***---

Abstract - Social media platforms have been substantially contributing to stoner- generated content. Similar social media data correspond of colorful themes bandied online and are associated with sentiments of the druggies. To catch up with the speed of streaming data at which it generates on social media platforms, it's pivotal to descry the motifs being bandied on social media platforms and dissect the sentiments of druggies towards those motifs in an online manner to make timely opinions. Sentiment analysis (SA) on social networks like Twitter and Facebook has numerous uses and has developed into a important tool for understanding stoner opinions. Sentiment analysis also assists in determining the feelings. In the proposed design, sentiment analysis is employed to identify reviews using an RNN. A machine learning methodology that provides a more accurate estimation of the sentiments than the presently employed styles. Delicacy is enhanced when used on a larger body of big data, where it'll demonstrate its significance

Key Words: Sentiment analysis, machine learning methodology, Deep Learning.

1. INTRODUCTION

Social networking's rapid expansion has altered how people engage with one another, impacting and ultimately changing their way of thinking their opinion on some issues. Therefore, the examination of content created by users' great approach to instantly detect Developments and perspectives on a topic [11]. Twitter, LinkedIn, Facebook and many other popular online social networks it's becoming more and more popular. In addition, numerous multimedia networks Flickr has also become popular in recent years. Many Such social networks are very content-rich and typically huge amount of content and link data available for analysis. Link data is Communication between units is generally represented by a social network diagram, whereas content data on the network comprises text, photos, and other multimedia data. In the context of social networks, the wealth of this network offers Unimaginable opportunities for info. Analysis. Extracting preferences is the science of analysing text. Understand the factors behind public opinion. Assessment of sentiment [15] is the examination of opinions. Examination of perspectives is another level of a ball game. Mood for an instance this looks at how everyone thinks (positively or negatively) about a particular topic. Examination of opinions to find out why individuals experience what they do. In this application, we present an application based on available information from twitter. Our system uses Bayesian logic to perform probabilistic inference network. Our approach will be useful in most of the domains like healthcare, tourism, political opinions, economic issues, etc. Using Twitter data collected via API, create a Bayesian algorithm that can be used to analyses new user-identified tweets.

In this, we demonstrate a technique for mining opinions using data from Twitter. Our approach uses Bayesian networks to carry out probabilistic logical reasoning. Our method can be used in many different areas, including health care, travel, political views, economic concerns, etc. Use the joint probability distribution to calculate the likelihood of getting from one position to the next. The combined distribution of probabilities, as defined by mathematics, is the likelihood that two or more occurrences will occur simultaneously. To build and train a Bayesian algorithm and embed it in an application that can be embedded into new Tweets detecting user's wishes, we used Twitter data that was gathered through APIs. For instance, every social media member has the ability to share images of locations they have visited

2. LITERATURE REVIEW

Everyone is eager to convey their political inclination of the aftermath in the social involvement among young people in particular [5]. But as intrigued by the voice of the people. Social media applications, where the many facets of preference of the people can be quickly extracted and taken, are widely acknowledged as the ideal outlet to this rising need of political activism. These websites are the impact of social media is becoming increasingly pronounced in shaping people's thoughts and actions. It is commonly acknowledged that the views held by the general public greatly influence the outcome of political parties and are a direct reflection of their governance. While increasing public feedback sharing has fostered accountability and raised awareness, it has also caused turmoil and uncertainty for many. This paper proposes a method to facilitate opinion mining through linguistic analysis and opinion classifiers. The study aims to streamline the process of sentiment analysis, using the popular microblogging platform, Twitter, to identify neutral, negative, and positive opinions related to political parties in Pakistan. A methodology is presented to categorise this data which is shown in figure 2.2 that pre-processes the Twitter raw data and contrasts two categorization strategies. Aiming to take a snapshot of the current political landscape will help Pakistani politicians feel more responsible, self-aware, and forward-thinking.

3 PROPOSED SYSTEM

In this section, we provide the details on the dataset used, the proposed architecture for the process of examining opinions from tweets. In this, we demonstrate a technique for mining opinions using data from Twitter. Our approach uses Bayesian networks to carry out probabilistic logical reasoning. Our method can be used in many different areas, including health care, travel, political views, economic concerns, etc.

Data Integration Tweepy is one of the most basic APIs we can use in python [2]. In order to authorize Twitterto access, we need Keys and Access Tokens. For quickly iterating through multiple object types, Tweepy provides the useful Cursor interface. By using this Cursor interface, we can extract tweets by the search key or by a Hashtag. It gives access to all tweets as they are published on Twitter as shown in figure 3.1. We search the particular tweets according to the tourism domain we need. are determined by the architecture.

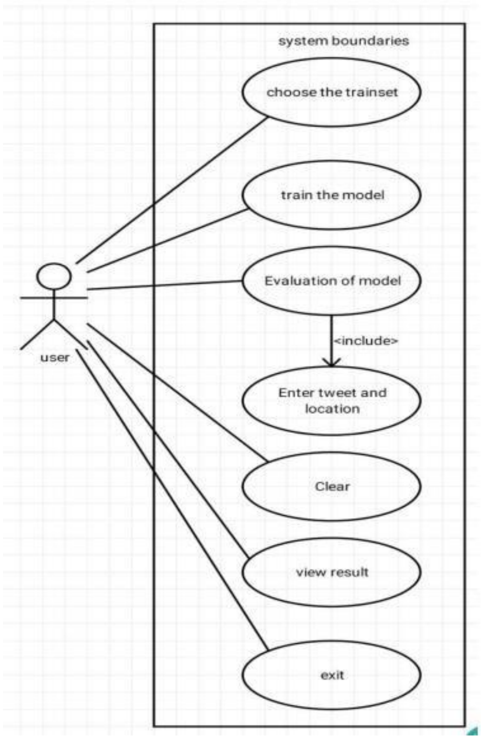

Figure 3.1 Use case diagram

Use case illustration an example of a behavioral diagram is the Unified Modelling Language

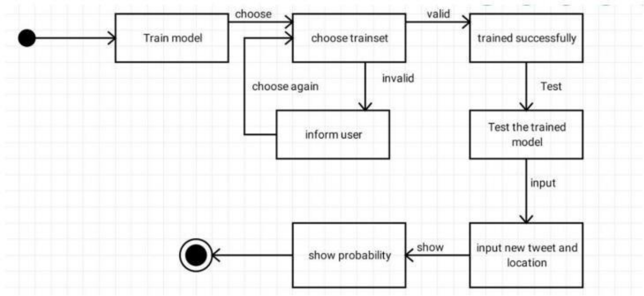

Figure 3.4 State chart diagram

4. EXPERIMENTAL RESULTS

Certainly, A state chart diagram, also known as a state machine or state chart diagram, is a representation of the states that an object may reach as well as the transitions between thosestates using the Unified Modelling Language (UML). It is shown in the above figure. In allforms of object-oriented programming (OOP), state diagrams.

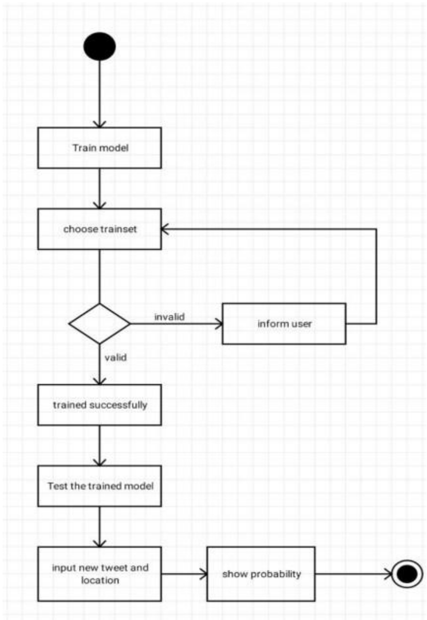

Figure 4.1 Activity Diagram

An activity diagram is a visual representation of how a specific job leads to the next, using a system operation to describe the process. The flow of control is indicated by connecting one activity to the next

4.1 Module DescriptionPre-Processing Module

Evaluation of sentiments and entity detection automatically produce independent of the provided training dataset for the parameters and conditions. The sentiment analysis technique uses the VADER tool, and its output is used to automatically generate up to two Random variables. Entity recognition is often used in the process of locating the independent variables created on the purpose of a given tweet when a unique value that makes an independent variable true if discovered on a tweet.

4.2 Sentiment Analysis: The VADER tool returns an emotion value. When the value reflects a positive emotion, the value of the parameter gets established, indicating that this variable is good and running. If the value indicates a negative sentiment or a negative integer, a randomly generated negative Sentiment is created, signaling that this variable is false otherwise. If VADER yields a value of zero, the positive Sentiment and negative Sentiment random variables are given the value Null.

This process involves the following steps

4.2.1. Processing: The tweets are pre-processed to remove noise, such as URLs, hashtags, and user mentions.

4.2.2. Feature extraction: Features such as n-grams, emoticons, and hashtags are extracted from the pre-processed tweets

4.2.3. Classification: The extracted features are used to classify each tweet into one of three sentiment categories: positive, negative, or neutral [10].

4.2.4. Bayesian inference: The sentiment classification results are used to update the hidden sentiment variables in the Bayesian network model using Incremental Learning.

4.3 Entity recognition: Entity recognition is a Identifying and extracting entities from text, such as names of people, places, and organizations, is a Natural Language Processing (NLP) problem. In the context of sentiment analysis on Twitter data Entity recognition can be useful for identifying the entities that are being discussed and their associated sentiments, which can provide additional context and improve the accuracy of the sentiment analysis The model that our framework creates identifies Twitter users who may be tourists visiting A place; as a result, the randomly selected user Location refers to users whoown tweets and have home locations that are not in A location.

Algorithm:

Input: Twitter data

Output: likelihood that a person will visit a location

4.3.1 Creation of Random Variables and Rules

Using the algorithm shown below in the figure 3.6, discrete random variables are automatically generated. Random variables are generated by the algorithm for sentiment analysis. The random variable user Location is first entered in list random Variables. The user's residence makes up this random variable. Additionally, the function creates the random variable positive Sentiment and returns 1 if the arrayof tweets contains a positive sentence-tweet.

```
procedure sentimentIdentificationWithVADER(in: arrayOfTweets; out:
randomVariables)
begin
 randomVariables := [“userLocation”];
 if (sentimentAnalizer(arrayOfTweets) = 1) then
 randomVariables := append(randomVariables, [ ”positiveSentiment”]);
 if (sentimentAnalizer(arrayOfTweets) = -1) then
 randomVariables := append(randomVariables, [ ”negativeSentiment”]);
end
```

Figure 4.3.1 creation of random variables and rules

The inputs for the functions prefix, infix, and postfix are a tweet's tweet Text and the value from the table's row Synonyms [field Value] field. If the row's field value is present at the start, middle, or end, they return true. The tweet's conclusion. In that scenario, the list random Variables is updated with the field's associated value

4.3.2 Create Random variables using E.R Method

Each rule is assigned a probability based on the frequency of each incident of evidence in the evidence set. The probability of rules is calculated in follows: Probability of a rule equals number Of Occurrence / size of Array of Tweets, where "number of Occurrence" denotes the quantity of times each piece of evidence appears in the twitter array and "size of Array of Tweets" denotes the size of the data inside the tweet is in the form of array. Therefore, the corresponding evidence set is used todetermine the probable cause of the modes

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

```
procedure createRandomVariables(in arrayOfTweets; out randomVariables)
begin
 for each tweetText ∈ arrayOfTweets do
 for each rowSynonyms ∈ synonymsCSVtable do
 for fieldValue := 1 to size_of_rowSynonyms do
 if (prefix(tweetText, rowSynonyms[fieldValue]) or
 infix(tweetText, rowSynonyms[fieldValue]) or
 postfix(tweetText, rowSynonyms[fieldValue])) then
 if fieldname ∌ randomVariables then
 randomVariables: = append (randomVariables, fieldname);
end
```

Figure 4.3.2 creation of random variables using E.R Method

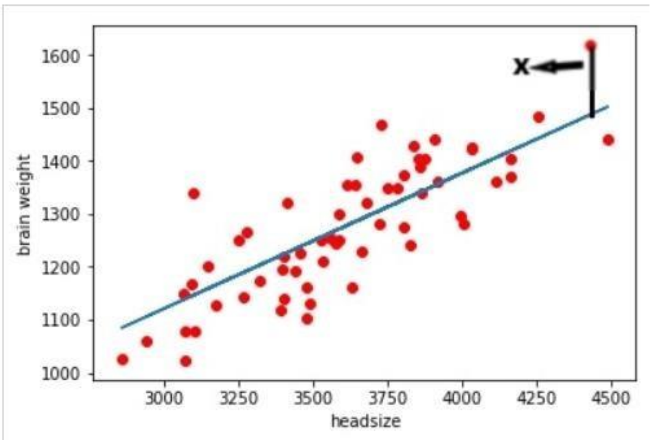

Figure 4.4 Evaluation Metrics

4.4 Mean Squared Error (MSE): MSD is another name for MSE. The MSE detail of action computes the average of the wrong squares (2). Additionally, it is non-negative, and a worththat is to say nearer to nothing advances the model's accuracy. When investigator of crime the foresaw worth and is the noticed worth, the mean regulated wrong in a mathematical model is deliberate utilizing the ability below

4.5 Root Mean Squared Error (RMSE): This gauges by virtue of what correct a machine intelligence model is. It is the average of the satisfied disagreements between the indicator and the real note, expected more exact [13]. It is completely complementary to the MAE measure; two together versification most frequently share otherwise-familiarize scores, that way that the tighter to nothing a calculation is, the more correct the model is (3)

4.6 Mean Absolute Error (MAE): is a measure of dissimilarity 'tween two unending variables.It is top-selling rhythmical for weighing the veracity of model. Additionally, MAE mainly stands by to the following rule: Actual Value - Predicted Value: Prediction Error [13]. Actual Value: is the worth found by way of scrutiny or calculation of the facts within reach. The noticed worth is another name for it (4). Value expected for one reversion model is famous as the average value. Every record in the dataset has allure prognosis mistake planned, and before we turn all negative mistake into a beneficial individual

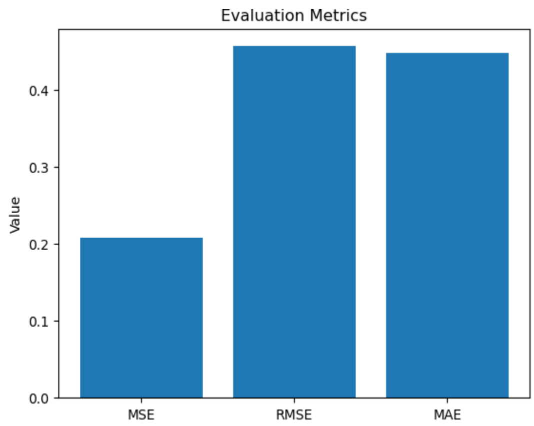

Figure 4.5 Graphical Representation of metrics

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

Where,
$\hat{y}$ − *predicted value of y*
$\bar{y}$ − *mean value of y*

The values which we calculated for out model are as follows:

- **RMSE:** 0.45
- **MAE:** 0.44
- **MSE:** 0.20

We also calculated the accuracy (5) and F1 Score (6) to make sure how efficient our model is running. Accuracy and F1 score are two commonly used performance evaluation metrics in machine learning. Accuracy measures the percentage of cases in the dataset that were correctly categorized out of all the instances. It is a simple measure that can be used when the classes in the dataset are balanced. However, it can be misleading when the classes are imbalanced, as it may give a high accuracy rate even if the model is only predicting the superior class.

$$\frac{TP + TN}{TP + FP + TN + FN}$$

F1 score is an average of the precision value and recall value, and it is often used in binaryclassification problems

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

The accuracy and F1 Score of our model are as follows:

ACCURACY: 93%

F1 SCORE: 0.96

5. CONCLUSION:

Our algorithm, which primarily focuses on determining if a user is likely to travel to a location, does opinion mining on social media data using probabilistic logical reasoning. This has real-world implications for the travel sector, particularly travel agents. We use entity identification and sentiment analysis to automate critical processes including producing random variables, formulating rules, and establishing evidence, which are crucial components of a Bayesian Network. The Prolog toolset is then used by the system to train the model. Our method is simply adaptable to analyze social media data on numerous themes, including Facebook, Instagram, and other sites in addition to Twitter. Our technology also supports incremental learning, allowing the model to keep becoming better overtime. Our system's automated chores are finished, and then it moves on to training the model with the Problog toolset. Our approach may easily be modified to different social media subjects in order to do opinion mining. We used Twitter in our technique, but it could be done to any other social networking application, including Facebook, and others. Our methodology encourages gradual learning so that the resulting model can be made better. Our strategy makes use of Twitter data as well as a probabilistic opinion mining tool. The actions that come next give an overview. First, the structure of our method uses the Twitter API to import a large number of Tweets. Following the automated handling of the imported Tweets, a set of random variables and untrained rules are generated. A Bayesian Network is then constructed utilizing a set of data, random variables, and untrained rule sets. The trained model can then be used to assess recent Tweets. The built-in model can also be incrementally retrained to increase its resilience. We choose to use tourism as the application domain because it is among the most widely used themes on social media. We were able to create this model with a 93% accuracy rate.

6. REFERENCES

1. Gepperth, Alexander, and Barbara Hammer. "Incremental learning algorithms and applications." In *European symposium on artificial neural networks (ESANN)*. 2016.

2. Gupta, Prasoon, Sanjay Kumar, Rajiv Ranjan Suman, and Vinay Kumar. "Sentiment analysis of lockdown in india during covid-19: A case study on twitter." IEEETransactions on Computational Social Systems , no. 4 (2020): 992-1002.

3. Cheng, Jie, G. Grainer, J. Kelly, D. A. Bell, and W. Lius. "Learning bayesian networks from data: An information-theory based approach, 2001." URL citeseer. ist. psu. edu/628344. html (2001).

4. Bucur, Cristian. "Using opinion mining techniques in tourism." Procedia Economics and Finance 23 (2015): 1666-1673.

5. Gull, Ratab, Umar Shoaib, Saba Rasheed, Washma Abid, and Beenish Zahoor. "Pre processing of twitter's data for opinion mining in political context." Procedia Computer Science 96 (2016): 1560-1570.

6. Rao, Praveen, Anas Katib, Charles Kamhoua, Kevin Kwiat, and Laurent Njilla."Probabilistic inference on twitter data to discover suspicious users and malicious content." In 2016 IEEE International Conference on Computer and Information Technology (CIT), pp. 407-414. IEEE, 2016.

7. Nozawa, Yuya, Masaki Endo, Yo Ehara, Masaharu Hirota, Syohei Yokoyama, and Hiroshi Ishikawa. "Inferring tourist behavior and purposes of a Twitter user." In Trends in Artificial Intelligence: PRICAI 2016 Workshops: PeHealth 2016, I3A 2016,AIED 2016, AI4T 2016, IWEC 2016, and RSAI 2016, Phuket, Thailand, August 22- 23, 2016, Revised Selected Papers 14, pp. 101-112. Springer International Publishing,2017

8. Prameswari, Puteri, Isti Surjandari, and Enrico Laoh. "Opinion mining from online reviews in Bali tourist area." In 2017 3rd International Conference on Science in Information Technology (ICSITech), pp. 226-230. IEEE, 2017.

9. Agarwal, Apoorv, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca J. Passonneau."Sentiment analysis of twitter data." In Proceedings of the workshop on language in social media (LSM 2011), pp. 30-38. 2011.

10. Kouloumpis, Efthymios, Theresa Wilson, and Johanna Moore. "Twitter sentiment analysis: The good the bad and the omg!." In Proceedings of the international AAAI conference on web and social media, vol. 5, no. 1, pp. 538-541. 2011.

11. Dave, Khyati, Surbhi Chandurkar, and Ashika Sinha. "Opinion mining from social networks." *International Journal of Computer Science and Network* 3, no. 6 (2014): 3-6.

12. De Raedt, Luc, Angelika Kimmig, and Hannu Toivonen. "Problog: A probabilistic prolog and its application in link discovery." In *IJCAI*, vol. 7, pp. 2462-2467. 2007.

13. Willmott, Cort J., and Kenji Matsuura. "Advantages of the mean absolute error (MAE)over the root mean square error (RMSE) in assessing average model performance." *Climate research* 30, no. 1 (2005): 79-82.

BIOGRAPHIES

G MAHESH CHALLARI "Received B.Tech in ECE from Andhra university 2108. Completed M.Tech in Communication Engineering & Signal Processing (ECE) from University College of Engineering Kakinada JNTUK 2022. Currently working as Assistant Professor in Department of Cse & coordinator R&D at Sree Dattha Institute of Engineering & Science "

K.VIJAY KUMAR "Received B.Tech in CSIT from kbr college jntuh 2012. Completed M.Tech in CSE from Nagole institute of Technology & Science Jntuh 2015. Currently working as Assistant Professor in Department of Cse at Sree Dattha Institute of Engineering & Science."