iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.45504

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

Twitter Sentiment Analysis Using NLTK and Machine Learning

Srilekha Vuppala[1], Spoorthy Singa[2], Sumanth Vasa[1], Kasi Bandla[4]
[1, 2, 3]*B.Tech Students,* [4]*Assistant Professor, Department of Electronics and Computer Engineering, Sreenidhi Institute of Science and Technology, Hyderabad –501301, India*

Abstract: ***In today's online social networks like twitter all people choose to express their opinions on social networking sites about the products or organizations, if any of the user has a good experience with any of the product or company, he/she will express their views which can be good reviews/opinion by seeing these opinion other users can know the quality of the product. On Internet, opinion mining which can be on sentiments or topic helps users to know the quality of any of the organizations or products, while developing new techniques to detect the sentiments from these opinions, all existing techniques that are used to discover are either Positive or Negative or Neutral sentiments from topics but this paper proposes 5 levels of sentiments detection such as High Positive, Moderate Positive, Neutral, High Negative and Moderate Negative. To detect sentiments, we are using four Ordinal Regression machine learning algorithms such as SoftMax, Decision Tree, Random Forest and also Support Vector Regression. For classification of tweets, we used NLTK, which cleans the tweets by removing special symbols, removing stop words, word stemming, etc. In this paper the authors have discussed how these algorithms are implemented on tweets and detect the sentiments***
Keywords: Machine learning, Opinion mining, Sentimental analysis, Social networks.

I. INTRODUCTION

What do most of the people of the current generation do when they want to express themselves, be it anonymously or individually? they log in to internet mainly via their social media platforms and post anything they feel or want to express. At this current era of digitalization, it has become common for people to express views through internet via social media, blogging, micro-blogging etc., there are various platforms and applications where people can choose to express their views about, and twitter is such an application. Twitter has become current most basic platform where people present their opinions in the form of tweets and also twitter generally is the world's largest microblogging social site with a huge user base of 330 million people tweeting around 500 million tweets every day who share their views and opinions through tweets, according to a statistics analysis by [2].

Hence here the author chose to conduct sentiment analysis on twitter. Sentiment analysis is frequently carried out at several levels, ranging from coarse to fine.

The main goal of coarse level analysis is to determine the sentiment score of the entire document, whereas fine-level analysis focuses on the attribute level. Between these two is sentiment analysis at the sentence level. Twitter sentiment analysis is being widely used in many areas, be it detecting sentiments during quarantine [3] or be it at polling and guessing who is going to win the elections at that period of time [4] and has most of the times successfully predicted the right outcome. In our project we are utilizing ML to tackle with twitter sentiment analysis. The general classification algorithms are focused on foreseeing nominal data labels. However, to rule for predicting categories or labels on an ordinal scale involves many patterns recognition issues. This type of problem is known as ordinal classification or ordinal regression.

In the further chapters of this work the authors discuss about the problem that is in the current project and the solutions they have provided to it, we also learn about its methodology using the algorithms such as SVM, Random Forest, Decision tree and shows which classification algorithm gives the better results and the analysis of the project.

II. LITERATURE REVIEW

Twitter sentiment analysis is a topic that has been researched through many years. Various researchers have been working on twitter from ages and have been issuing their self-found researches. The researchers have been using numerous sentiment analysis techniques for improvising the outcomes of the classifications. The first basic system was classifying the tweets into positive and negative using sentiment recognition based on textual data as NLP and machine learning using maximum entropy [1].

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

Their work is likewise useful in this examination as the feeling investigation strategies they have utilized, highlight determination methods, different pre-handling steps they have utilized is dealt with in this exploration. This exploration primarily centres around directed approach for opinion investigation task and has reviewed investigates both for twitter and non-twitter information and furthermore for both regulated and vocabulary-based approaches for better explanation and comprehension of the subject picked.

Further most of the researchers have used KBA (Knowledge based approach) that contributes considerably to analyze tweet sentiments/emotions. Knowledge-based coaching is an approach that involves adapting theories, knowledge, and traditions.[6]. Also, there were other methods utilized like SVM and KNN based hybrid classification model which is presented to process the tweet features and then to identify the unseen sentiments from these tweets.[5].

For further better classification researchers started using naive bayes algorithm along with SVM [9] and checked and improved the accuracy rates accordingly which were 81 and 67% simultaneously. This paper classified the tweets into 3 different classifications as positive, negative and neutral. This paper [8] provides two different ML techniques called naive bayes and SVM to understand and research sentiment analysis. It basically uses NLP and ML to identify the polarity of the tweets.

As Sentimental analysis has been a growing bigger at the area of NLP with researches varying from document level classification (Pang and Lee 2008) for knowing the polarity of phrases and words (e.g., (Hatzivassiloglou and McKeown 1997; Esuli and Sebastiani 2006)). As twitter has a character restriction of 280 characters including spaces, hashtags, symbols etc.., classifying the sentiments of the tweets comes under sentence level sentiment analysis (e.g., (Yu and Hatzivassiloglou 2003; Kim and Hovy 2004)); Although, with the informal and specialized characters that are utilized in tweets along with the essence of microblogging domain making twitter sentiment analysing a challenging work. Yet, it's still unclear that how effectively features and approaches based on more structured data will be translating to microblogging.

Challenges in understanding the sentiments that are expressed. In addition to that there is a need for automatic techniques that require large datasets of annotated posts or lexical databases where words are associated to sentiment values.

The authors have examined and solved the equations of the selected tweets and later on recognized the tweet ssentiment by using Natural Language Processing Toolkit which then classify the tweets into 5 different levels like - Highly positive, Moderate Positive, Neutral, Moderate Negative and Highly negative.

After performing these required actions, python programming Language provides libraries which helps for the implementation of the model proposed.

To detect sentiments, in this paper the authors have used four Ordinal Regression ML algorithms such as SoftMax, Decision Tree, Random Forest and also SVM.

Ordinal Regression is the classifier used with many independent variables to predict class of given data. In this paper, tweets took from twitter as input, applied to the classifier which predicts sentiment by using all independent words.

III. METHODOLOGY

In this module we see what classifications and algorithms the authors have used to classify the tweets based on their polarity, how the system functions and the 5 types of classifications.

A. Support Vector Machine

So, what precisely is Support Vector Machine (SVM)? We'll begin by grasping SVM in straightforward terms. Suppose we have a plot of two-mark classes as displayed in the figure beneath: Fig 3.1.

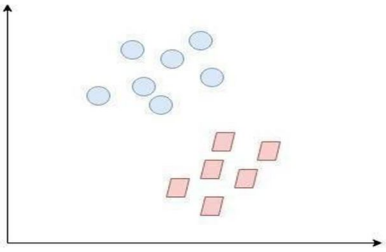


FIGURE 3.1, Graph with two classes

Might you at any point choose what the isolating line will be?

You could have thought of this: Fig 3.2.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

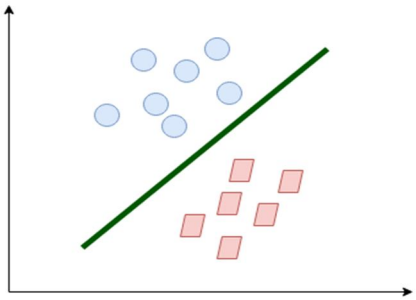


FIGURE 3.2, Simple Class Separation

The line decently isolates the classes. This is the very thing that SVM basically does straight forward class division. Presently, what is the information was this way: Fig 3.3. Here, we don't have a straightforward line isolating these two classes. So, we'll expand our aspect and present another aspect along the z-pivot. We can now isolate these two classes: Fig 3.4.

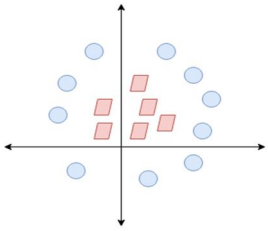


FIGURE 3.3 SVM Dimension Extension

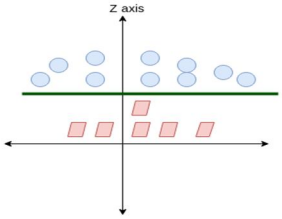


FIGURE 3.4 SVM Dimensions

At the point when we change this line back to the first plane, it guides to the roundabout limit as I've displayed here:

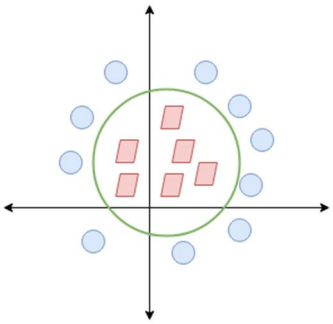


FIGURE 3.5, SVM Hyperplane

This is the very thing SVM really does ! It searches for a line or hyperplane (in multi-faceted space) that separates these two classes. The new point is then characterized in light of whether it is on the positive or negative side of the, still up in the air by the classes to be anticipated.

B. Random Forest

We should initially find out about the gathering method before we can understand how the arbitrary woodland functions. Various models are consolidated in a group. Accordingly, instead of utilizing a solitary model to make expectations, an assortment of models is utilized.

This algorithm uses two types of methods:

1) *Bagging:* It makes an alternate preparation subset from test preparing information with substitution and the last result depends on greater part casting a ballot. For instance, Random Forest.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

2) *Boosting:* It converts weak learners into strong learners by building sequential models with the best accuracy as the final model. For example, ADA BOOST, XG BOOST

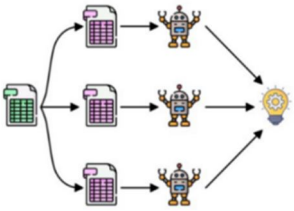


FIGURE 3.6 Bagging (Parallel)

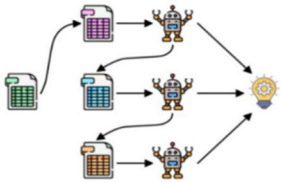


FIGURE 3.7 Boosting (Sequential)

a) Decision Tree

Decision tree is the extraordinary and well-known algorithm for order and expectation. A Decision tree is a flowchart like tree structure, where each inside hub indicates a test on a characteristic, each branch addresses a result of the test, and each terminal hub holds a class name.

Decision tree is able to create comprehensible rules. It can execute classifications without using much computation. Decision trees are able to manage both continuous and categorical variables. It provides a clear indication of which fields are considered to be of much importance for prediction or classification.

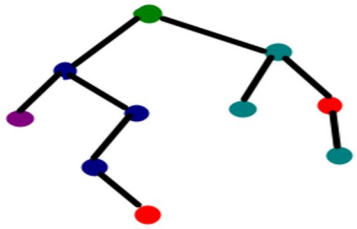


FIGURE 3.7 Decision tree

b) Summary Of Results

Algorithm	Accuracy
SVM	71.35
Random Forest	88.9
Decision Tree	94.5

TABLE 1: Summary of Results

As the table shows, when the processing, analysis was done on the bigger dataset, the accuracy scaled up to a great extent. Support Vector Machine algorithm scaled up to 71.35 and Random Forest scaled up to 88.9 percent and Decision Tree scaled up to 94.5%. The best result tested thus far, was obtained when Decision Tree is used on the data, gave an accuracy of 94.5 percent.

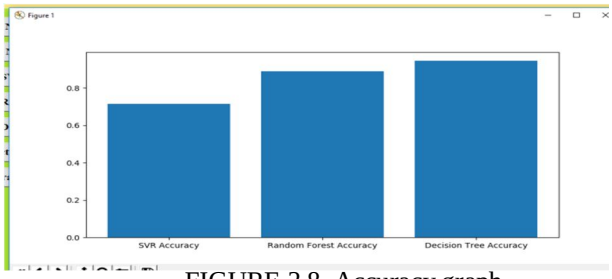


FIGURE 3.8, Accuracy graph

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

IV. CONCLUSION

In order to build an opinion on anything and form a foundation for it, Twitter analysis is mostly used to identify the sentiment type of various tweets and scenarios. With five classifications of tweets—highly positive, moderately positive, neutral, moderately negative, and highly negative—the authors of this paper, presented sentimental analysis of tweets taken from Twitter using different ML algorithms like SVM, Random Forest, and Decision Tree using NLTK, achieved accuracy of 71%, 89%, and 94%, respectively.

Our project can also be further done in other complex projects with different tasks, such as online calculators, it can simply take a picture of the equation and calculate its solution for the equation without extra input or human intelligence. This project has a number of advantages, mainly time saving when solving an equation, which will certainly be useful for many fields whose schedules come with time consuming solving and analyzing the equations.

REFERENCES

[1] Shaunak Joshi, Deepali Deshpande "Twitter Sentiment Analysis System" International Journal of Computer Applications June 2018.
[2] Shea M Lemley, Jeffrey D Klausner, Sean D Young, "Comparing Web-Based Platforms for Promoting HIV Self-Testing and Pre-Exposure Prophylaxis" (2020): October.
[3] Mohammad Abu Kausar, Arockiasamy Soosaimanickam, Mohammad Nasar, "Public Sentiment Analysis on Twitter Data during COVID-19 Outbreak" IJACSA, 2021.
[4] AnkitaSharmaa UdayanGhoseb "Sentimental Analysis of Twitter Data with respect to General Elections in India" ICITETM2020.
[5] Ankita Gupta, Jyotika Pruthi, Neha Sahu, "Sentiment Analysis of Tweets using Machine Learning Approach" IJCSMC April 2017
[6] Riya Suchdev, Pallavi Kotkar, Rahul Ravindran, Rahul Ravindran "Twitter Sentiment Analysis using Machine Learning and Knowledge-based Approach" October 2014
[7] Shubham Kumar, Akhilesh Kumar Srivastava, Yaman Soni, Utkarsh Tyaghi and Nikhil Kumar Singh, "Twitter Sentiment Analysis: A Novel Machine Learning Approach" International Journal of Control and Automation (2020).
[8] Khushboo Gajbhiye (&) and Neetesh Gupta, "Real Time Twitter Sentiment Analysisfor Product Reviews Using Naive BayesClassifier" Springer Nature Switzerland AG 2020
[9] Vishal A. Kharde, S.S. Sonawane "Sentiment Analysis of twitter Data: A Survey of Techniques" IJCA April 2016
[10] Sakshi Koli, Ram Narayan, "Review Paper on Sentiment Analysis Technique by Different Machine Learning Approach" IJCSE NOV 2019
[11] https://www.kaggle.com/datasets/himanshupoddar/zomato-bangalore-restaurants?resource=download.
[12] Abinash Tripathya, Ankit Agrawalb, Santanu Kumar Rathc, "Classification of Sentimental Reviews Using Machine Learning Technique" ICRTC-2015.

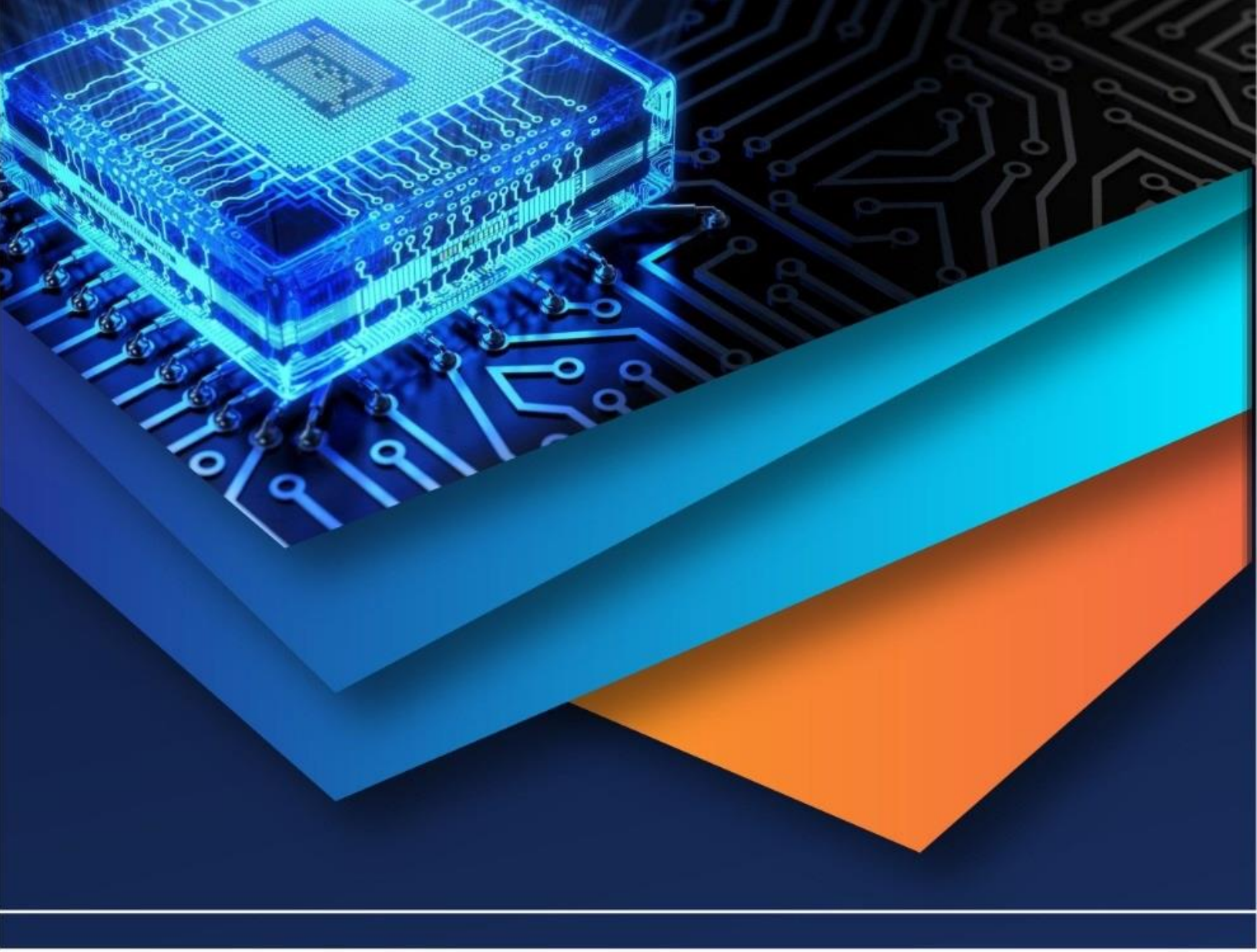
Srilekha Vuppala pursuing Electronics and Computer Engineering at Sreenidhi Institute of Science and Technology, Telangana, India. Also, she is working as an intern at Manjeera Digital Systems Pvt. Ltd., Telangana, India, where she is involved in Artificial intelligence and Machine learning domain.

Spoorthy Singa pursuing the B.Tech. degree in Electronics and Computer engineering from Sreenidhi Institute of Science and Technology, Telangana, India. Also, she is working with a well known Indian company Tata Consultancy services Limited(TCSL) as an assistant software analyst, Telangana, India.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

Sumanth Vasa pursuing the B.Tech. degree in Electronics and Computer engineering from Sreenidhi Institute of Science and Technology, Telangana, India. Also learning Python and Artificial intelligence course for his dream company.

Kasi Bandla, graduated with B.Tech (ECE) from BEC in 2006, received M.Tech (Microelectronics and VLSI Design) degree from SGSITS in 2009. Presently working as Assistant Professor in the Dept. of Electronics and Computer Engineering, SNIST, Hyderabad. Also, research scholar at BITS-Pilani, KK Birla Goa Campus, Goa. His research interests are Microelectronics, Embedded Systems and IoT Applications.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY