Performance Prediction of Product/Person

1st Devesh Bhangale
B.E., IT department
RAIT, Mumbai University
India
bhangalecr7@gmail.com

2nd Sneha Poojary
B.E., IT department
RAIT, Mumbai University
India
snehapoojary1202200@gmail.com

3rd Sameer Ahire
B.E., IT department
RAIT, Mumbai University
India
sameerahire6399@gmail.com

***Abstract*—Over the previous decade people have experienced exponential boom in the use of online resources, in particular social media and microblogging web sites such as Twitter, Facebook, Instagram and YouTube. Many businesses and agencies have identified these sources as a rich mine of marketing knowledge. On such platforms, massive quantities of records are produced (e.g.: 5000 tweets per $2d$ on twitter), this representing an chance for companies to check their social impact and people opinions towards their products, and even frequent people can additionally discover out what is a performance of a certain product or the overall performance of a particular political personality. In this project, we fetch the given number of tweets from user and classify it as Positive, Negative and Neutral by the usage of supervised machine learning approach. In this method we're analyzing the Polarity and Subjectivity of the tweets and then later we're using NLP to Classify the raw records into records body which gets rid of the undesirable words from each of the tweets. Neutral words like 'as, the, of' are eliminated from the tweets. Using NLP, we get better results of the tweets, later we classify the tweets using classifying algorithms like Random Forest Classifier, Decision Tree Classifier, Logistic Regression, Support Vector Classifier. Later it compares the result of tweets which had been analyzed before processing into NLP. We are also using Data Visualization for phrase frequencies, and for displaying a pie or bar chart of variety of positive, negative and impartial tweets.**

***Index Terms*—Twitter, sentiment analysis, Machine Learning, natural language processing.**

I. Introduction

Sentiment analysis is the automated process of identifying and classifying subjective information in text data. This may be an opinion, a decision, or a feeling about a specific subject , individual or product feature. 'Polarity detection' is the most common form of sentiment analysis and involves classifying comments as positive, negative or neutral. The Natural Language Processing (NLP) is used in the sentimental analysis to make sense of human language, and machine learning for automatically delivering the accurate results. Sentimental Classification and analysis is performed in python nltk module, Python has a special NLTK module to perform tasks in natural language processing. In our project, we are going to predict the performance of the product in the market, like what humans feel about the product (positive opinions or negative reviews) that has been launched hence helping the corporations to be aware of how their product is performing in the market and what they can improve/ alternate in their upcoming products. Our mission will now not solely help the businesses to get the performance of their product however additionally to the common people who prefer to comprehend how the product is performing so that it can help them while shopping for product, our module additionally helps to get the positive, negative and neutral prediction on the political party or a individual so that people can be aware of the performance of the person. The overall performance of the product/person will be anticipated with the help of the real-time facts that we get the usage of the Twitter API. The classification that we are going to used are Support vector Classifier, Random Forest, Decision Tree, Logistic Regression.

Identify applicable funding agency here. If none, delete this.

II. Problem Statement

Twitter, with over 313 million 1 monthly energetic users and over five hundred million tweets per day, has now turn out to be a goldmine for businesses and people who have a sturdy social, political, or economic interest in maintaining and bettering their product and client base. Twitter sentiment evaluation provides companies , organizations and even the common people the capability to monitor public emotions toward the merchandise and activities related to them in real-time. This venture helps to bridge the hole between the business enterprise and its customers. This technique can also be used to analyze the performance of politicians, sports person or an athlete over a duration of time.

III. Objective

- Traditionally companies used behavior questionnaires and surveys to achieve comments and perception into how customers felt about their products. These ordinary strategies were frequently extraordinarily time ingesting and costly and did not usually return the outcomes that the organizations have been searching for due to environmental factors and poorly designed surveys. Natural language processing and sentiment analysis are taking part in an increasingly vital function in making educated choices on marketing strategies and giving valuable comments on merchandise and services. In making informed decisions

on advertisement and marketing campaigns and getting useful input on products and services, natural language processing and emotion analysis play an increasingly necessary role. Thus, the ultimate product/man or woman performance is now not just for the enterprises but also for the common people before purchasing some product or when knowing the opinion of the person.

- This project will be useful to the companies , political parties as properly as to the common people. It will be useful to political party for reviewing about the software that they are going to do or the software that they have performed. Similarly companies can additionally can get evaluation about their new product on newly launched hardware or software.
- To put in force an algorithm for computerized classification of the textual content into positive, bad and neutral. Graphical representation of the sentiments in the shape of the pie-chart.

IV. Methodology

In our project we are going to predict the overall performance of the product in the market, like what people sense about the product(positive critiques or terrible reviews) that has been launched accordingly assisting the corporations to understand how their product is performing in the market and what they can improve or change in their upcoming products. Our task will now not solely assist the organizations to get the performance of their product however also to the frequent humans who choose to comprehend the how the product is performing so that it can assist them while shopping for the product, Our module additionally helps to get the positive, bad and neutral prediction on the political party or a character so that human beings can know the overall performance of the person. The overall performance of the product/person will be predicted with the help of the real time facts that we get using the Twitter API. The classification that we are going to used are Support vector Classifier, Random Forest, Decision Tree, Logistic Regression. For this project, we are using the Android framework; we plan to create an application as common people can access it. The App offers functions to obtain the product/performance person's prediction. This app can enable businesses to understand how their brand performs in the industry. Even ordinary consumers may either understand the individual's success or would get a public opinion of a product before purchasing it. The app displays the proportion of the total positive, negative and neutral tweets in the form of a graph, making it simple for businesses and even for ordinary people to understand the product/results.

A. *Use of Flask Framework*

Flask is a web framework. This potential flask offers you with tools, libraries and applied sciences that allow you to build a android application. This Flask provides an API that can be used as a back end for android apps to run the python file in android applications.

B. *Website*

The internet site is been developed for the laptop use which are often used in the agencies and the organizations. The website is developed with the help of the $Bootstrap$ and CSS is used for the designing of the page. The first page is the login page. If the consumer is new the person can register himself/herself in the signing page. After the login system the user enter the home web page, the home web page consists of the two columns, one the topic column the place the person enters the theme on which he/she wants to find the analysis. Second column is for the number of tweets where the person enters the quantity of tweets on which the evaluation has to be performed. After entering a topic and number of tweets the person clicks on the analyse button. This lands on the end result web page, right here the first end result web page consists of the word cloud and the bar plan which shows the range of tweets that are high-quality and negative. Also, the wide variety of the frequency of the word existing in the tweets. The fourth and the final page consists of the effects of the classification algorithms, it consists of the accuracy percentages and the $f1$ score. Thus by means of searching at these outcomes the consumer can decide and take the respective selection on the particular products/ person.

V. Literature Survey

[1] A twitter evaluation of Republic of Indonesia presidential candidates conducted in 2019 for the sentimental analysis, using the python programming language. Countless steps were taken for the behavior of this sentimental analysis ,which are the gathering of the records , using the python libraries, text processing ,testing education data and the text classification using the Naive Bayes method. They did the textual content processing from statistics got and use Naive Bayes method to predict the class. Later, this was compared with different strategies such as SVM and KNN. Then it is classified by using two instructions namely positive and negative . From the experimental results , it can be seen that the Naive Bayes technique has a better accuracy degree (i.e. 80.90%) in contrast to using other methods, such as KNN which only has an accuracy rate of 75.58% and an accuracy charge using SVM which is 63.99%.

[2] To advance a practical classifier which can automatically classify the sentiments of the unknown tweet. For the classification of the label accurately some methods were proposed. There are two methods: One technique is recognised as a sentiment classification algorithm (SCA) based on the K-nearest neighbour (KNN) and the other is primarily based on the support vector machine (SVM). Then it is tried to find out the positive and negative emotions on Twitter data. The work has been done with the help of simpler model and have also used the SVM (Support Vector Machine) machine learning algorithm. Along with the KNN classification and calculated the accuracy of all algorithms. After result it is seen that the Sentiment classification algorithm (SCA) works better than SVM. The accuracy of the result can be betters

by including the following :working with emotional tweets, working with multiple tweets, calculating accuracy, and evaluating performance.

[3] The records set consists of the tweets that are of 6 predominant US Airlines and multi-class sentiment evaluation is carried out. The strategy begins with pre-processing techniques used to clean the tweets and then representing these tweets as vectors the utilization of a deep learning concept (Doc 2 vec) to do a phrase-level analysis. The 7 distinct classification techniques had been used, they are deep learning concept (Doc 2 vec) to do a phrase-level analysis. The records is divided into $80 : 20$ percentage 80% for the education and the rest 20% for the testing. The end result of the take a look at set is in the structure of positive, negative, and neutral. In this paper, they have in distinction greater than a few ordinary classification strategies and study their accuracy. The classification techniques used consist of ensemble approaches such as AdaBoost it combines the end result of the exclusive classifier into one sturdy classifier and it affords an accuracy of 84.5%. The future work consists of the usage of a large data set.

[4] This examination plans to play out an itemized evaluation investigation of tweets structured on ordinal relapse making use of Artificial Intelligence procedures. The strategy that has been used as, first the pre-processing of the tweets and making use of element extraction strategies that supply the educated component. At that point, under a few classes, these highlights scoring and adjusting. Multi-nominal techniques like relapse, Support Vector Regression , Decision Trees, and Random Forest is used for the calculations which are used for the conclusion investigation grouping in the proposed system. For the utilization of this framework, a Twitter dataset used to be openly made available by using the NLTK corpora property used to be utilized ordering tweets into a few ordinal classes using AI classifiers. Classifiers, for example, Multinomial strategic relapse, Support vector relapse, Decision Trees, and Random Forest, are utilized in this examination.

[5] The experiment shows that the accuracy and the $F1$-score of the Twitter sentiment classification . The classifier was improved by using the pre-processing methods of replacing negation and expanding acronyms. But by removing the URLs, getting rid of numbers, or stop words no modifications was seen. When more than few pre-processing strategies were applied the Naive Bayes and Random Forest classifiers were extra sensitive than Logistic Regression and support vector machine. Experimental consequences indicate that the overall performance of the classifiers was affected by removing the URLs, the elimination of give up words, and the removal of numbers ; Furthermore the accuracy of the classifier was enhanced by changing the negation and expanding acronyms. In future the system can be enhanced by the use of exceptional stop-lists and acronym dictionaries and will look into the reasons for the fluctuation of sentiment classification performance the use of specific classifiers on a number datasets.

VI. Techniques:

A. *Software Model*

This chapter showcases the flow of our approach towards developing the software keeping in mind all the major steps of the (SDLC) Software Development Life Cycle.

B. *Software Development Life cycle*

Software Development Life Cycle is a method accompanied through all the software program manufacturing agencies to sketch for a best technique to manufacture software. This cycle includes a best graph describing how to develop, model, preserve and alter/enhance a unique software program and distinct levels of its manufacturing process. This helps in enhancing the quality of the software program and the ordinary improvement phase.

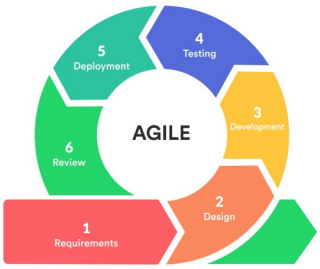

Fig. 1. Agile Methodology

1) **Communication**
   In this phase there must me proper communication between the stakeholders about the developments of the project. Properly explaining the plans and further process to the stakeholders is vital as they must also be aware of the project details.
2) **Planning**
   Any complicated journey can be simplified by proper planing and creating a map of the details which has to be followed in order to achieve the final goal. The planning involves all the software engineering works that are required to be conducted, all the risks involved in the process and give a proper work schedule to complete the task.
3) **Modelling**
   Modelling is a very crucial part wherein you decide the model of the project and a sketch of the thing so that you can understand the big picture and understand how the final product is going to look at its completion phase.
4) **Developing**
   The actual development of the project starts at this phase of the SDLC where the product is built. The programming code is written as per the project requirements. If

the design and modelling part is completed well then generating the code is a hassle free job.

5) **Testing**
This stage of SDLC is the subset of all the stages. Testing refers to the point where the project is tested with all the test cases and all the possible states where a bug can arise. This phase is very crucial as it helps in pointing the problems that may arise to a user. The bugs and issues are then reported, tracked and fixed until the product reaches the quality standards.

6) **Deployment**
In this phase the product is launched in the market and is made available in the public domain to all the users. The users are free to use the product and write there reviews and about it and give a feedback to the manufacturer. This is final step of the SDLC to launch the final product in the market.

C. Machine Learning Techniques

There are two basic ways of analyzing the sentiments from tweets, namely, Lexicon based and machine learning based approach. We have used the machine learning based approach for analyzing the data. The machine learning model is trained using a specific training data with known outputs. The several machine learning algorithms used are Support Vector Classifier, Decision Based Classifier, Logistic Regression and Random Forest Classifier.

1) **SUPPORT VECTOR CLASSIFIER:**
Support Vector Machines (SVM) are popularly and widely used for classification problems in machine learning. Support vector Regression comes under supervised machine learning algorithm which support the detection of classification, outliers and regression which are helpful for statistical theory of learning. The objective of the SVM algorithm is to build the best line or decision borderline that can divide n-dimensional space in group so that we can easily put the new data point in the correct division in the aftertime. This best decision borderline is called a hyperplane. SVM chooses the excess points or vectors that aid for building the hyperplane. Such excess cases are called as support vectors, and hence algorithm is Support Vector Machine.

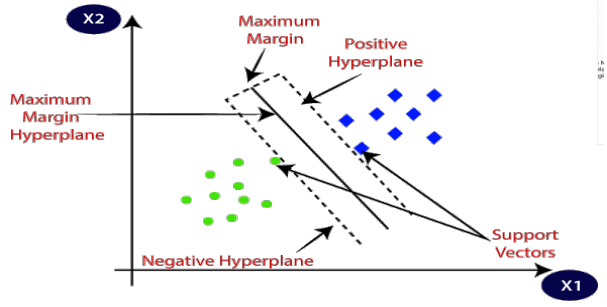

Fig. 2. SVM

2) **DECISION TREE CLASSIFIER:**
The classification technique is a systematic approach to build classification models from an input dat set. For example, decision tree classifiers, rule-based classifiers, neural networks, support vector machines, and naive Bayes classifiers are different technique to solve a classification problem. The decision tree classifiers organized a series of test questions and conditions in a tree structure.
ID3 algorithm use entropy to determine the correlation of a sample. If the sample is absolutely alike the entropy is zero. If the sample is an uniformly branched it has entropy of one
To frame a decision tree, we need to determine two types of entropy using frequency table.

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Fig. 3. 1st Type of Entropy

$$E(T, X) = \sum_{c \in X} P(c) E(c)$$

Fig. 4. 2nd Type of Entropy

The information gain is placed on the reduction in entropy after a dataset is split on an attribute. Build up a decision tree is all about finding attribute that returns the highest information gain

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

Fig. 5. Information Gain

3) **LOGISTIC REGRESSION:**
Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Some of the examples of classification problems are Email spam or not spam, Online transactions Fraud or not

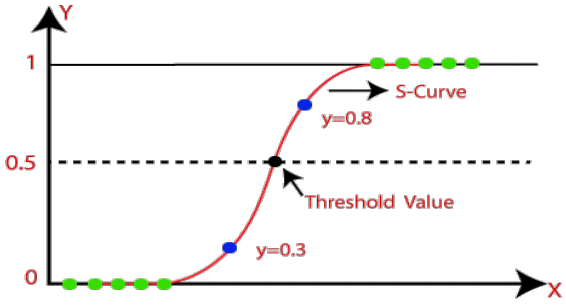

Fig. 6. logistic function (Sigmoid Function)

Fraud, Tumor Malignant or Benign. Logistic regression transforms its output using the logistic Sigmund function to return a probability value.

It is a model that in its basic form makes use of the logistic functions. It is the go to method for binary classification sums.It gives the probabilistic values which lie between 0 and 1. In Logistic regression, in lieu of fitting a regression line, we fit an ”S” shape logistic function, which will predict two maximum values i.e 0 or 1.

4) **RANDOM FOREST CLASSIFIER:** Random Forest algorithm is a supervised classification algorithm. We can see it from its name, which is to create a forest by some way and make it random. There is a direct relationship between the number of trees in the forest and the results it can get: the larger the number of trees, the more accurate the result. The random forest uses the combination of trees that can be used to analyze the class label depending on the categorical variable for a specified data.

VII. RESULT AND DISCUSSION:

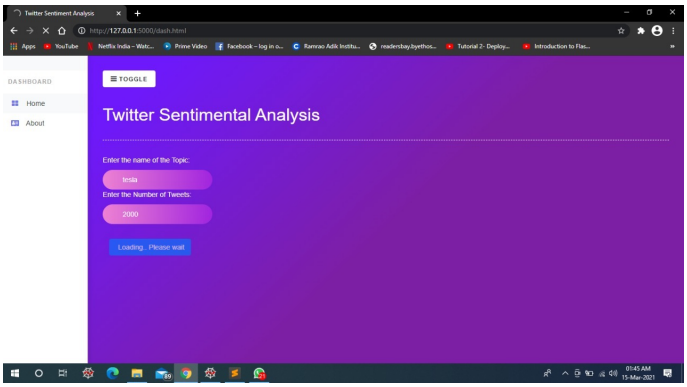

Fig. 7. Home Page

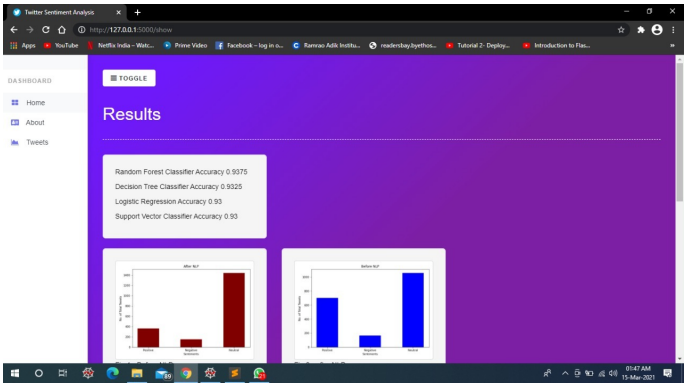

Fig. 8. Result Page 1

Experimental results point out that Support Vector Regression and Random Forest have an nearly comparable accuracy, which is higher than that of the naive Bayes and the logistic regression. However, the Decision Tree gives the highest

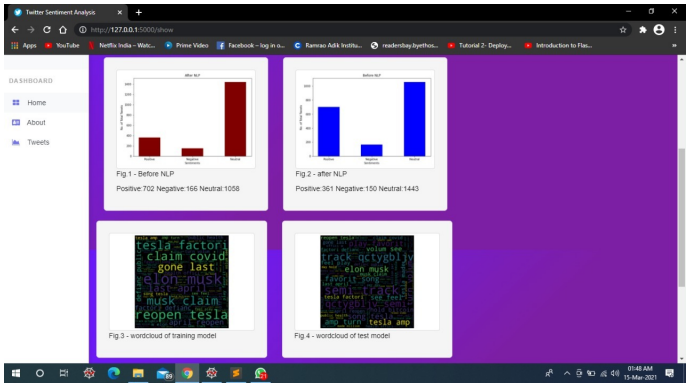

Fig. 9. Result Page 2

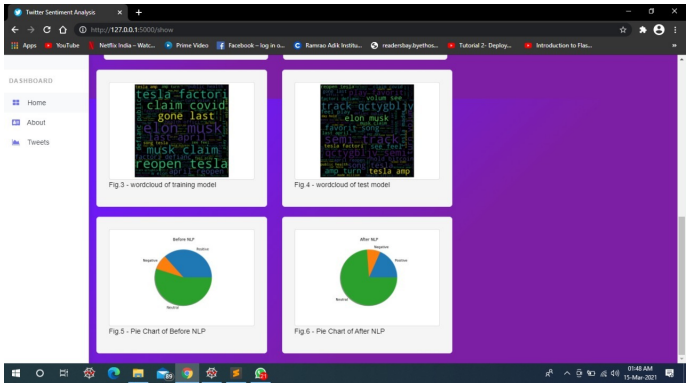

Fig. 10. Result Page 3

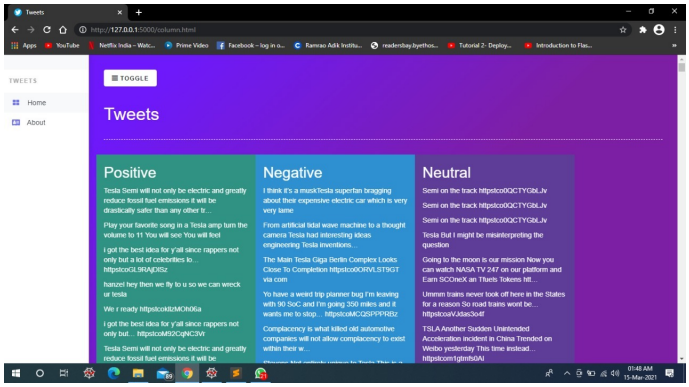

Fig. 11. Result Page 3

accuracy that is from 85% to 97%Experimental outcomes concluded that the Random forest and the Decision tree gives the best outcome as compared to others even if the results vary depending upon the the number of the tweets the user has entered. The result that is obtained from the classification process is stored in the file, later the data/result from the file is retrieved and shown in the website and if the android user the the result is showed in the app. The result is in the form of the bar layout, word cloud and the fundamental end result where the proportion of the accuracy is shown is introduced in the shape of the pie chart(the evaluation of how many positive,

negative and neutral tweets are present).

Obtained Accuracy

Accuracy	SVC	DTC	RFC	LR
Training	0.99	0.99	0.99	0.99
Validation	0.93	0.9325	0.9375	0.93
F1-Score	0.93	0.9325	0.9375	0.93

Note: Obtained Ac-curacies varies in every dataset of tweets which are fetched from twitter.

The word cloud tasks the frequency of the the phrases that are used in the tweets. The phrase frequency of both the tremendous and the terrible tweets are shown in the shape of the wordcloud. By looking at the the end result the groups or even the frequent people can get to comprehend about the performance of the political parties, character or the products.

VIII. CONCLUSION

Using sentiment evaluation of the Twitter tweets alternatives can be observed of a unique product, person, or community. Opinions can be positive, negative, or neutral relying upon the context of the tweets. Sentiment evaluation can be beneficial for the businesses for inspecting the overall performance of their merchandise in the public area and get the proper person evaluations out of it and this can be in addition used for further enhancements in upcoming products. The identical can be used to perform evaluation of upcoming elections, getting film reviews, and even overall performance of a individual over a duration of time. In this project, we are attempting to provide a easy and convenient way for the customers to apprehend the assessment of the merchandise that have these days launched. Along with that, the person will get to recognize the nice and bad opinions of the people. The employer or the businesses will understand how nicely their merchandise are performing in the market and the developments/ upgrades they have to make in their in addition updates. For example, the consumer right here has given the theme as tesla and the quantity of the tweets is 2000. Here we can see that we are the usage of 4 one of a kind algorithms to function the classification, the four exclusive algorithms are Random Forest Classifier, Decision Tree Classifier, Logistic Regression, Support Vector Classifier. We are using Data Visualization for phrase frequencies, and for exhibiting a pie or bar chart of a range of positive, negative, and independent tweets. From the figures below, we get that the random forest offers the great end result in contrast to all different algorithms. The Random forest offers the accuracy of 94% as the others provide much less accuracy percentage. The outcomes differ in accordance to the subjects and the range of tweets entered via the user.

REFERENCES

[1] Huma Parveen , Shikha Pandey(2019): *Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler : Twitter*.

[2] Mohammad Rezwanul,Ahmad Ali,Anika Rahman (2017):*Sentiment Analysis on Twitter Data using KNN and SVM.*

[3] Rane, A, Kumar, A. (2018): *Sentiment Classification System of Twitter Data for US Airline Service Analysis*.

[4] Saad, S. E., Yang, J. (2019): *Twitter Sentiment Analysis based on ordinal Regression.*

[5] Zhao Jianqiang ; Gui Xiaolin (2017): Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis.

[6] G.Kavitha , B.Saveen and Nomaan Imtiaz:**Discovering Public Opinions by Performing Sentimental Analysis on Real Time Twitter Data.**

[7] https://www.aclweb.org/anthology/W15-2914.pdf