PBL – II

Project Report

**Twitter Bot Analysis**

**Group Members** - 68 - Mohit Wani

61 - Aditya Shirke

57 - Aditya Sanas

58 - Ajinkya Sanas

143 - Sharad Sawant

**Assignment No. 1**

**Title of Assignment :-** Introduction and the need of the project .

**Theory :-**

One of the popular social media apps today is Twitter. It was launched in 2006 by Jack Dorsey . Twitter is a micro-blog service in which the users can send short messages, which then called tweets . Each tweet is limited to 140-character only. However, the simplicity and capacity to send tweets as often as possible become the additional value from this application. According to statistics in 2018, there have been 326 million Twitter active users each month, with an average of 500 million tweets sent every day .

Bots or automated programs are growing in popularity

alongside Twitter's popularity. They are created using the

Twitter Application Programming Interface (API). A study by

the United States Commission on Exchange and security in

2016 found at least 8.5% of active Twitter users were bots .

Bots have various objectives. It is undeniable that some bots have benefits in disseminating weather or earthquake information . However, many malicious bots exist too. They harm by broadcasting malware links , disrupting other users, broadcasting terrorist propaganda, spam, news lies, making hoaxes, and doing political campaigns. A massive tweet volume is capable of polluting users' timeline, changing user

perception, damaging user confidence, affecting the stock market, and even being able to undermine social order.

Therefore, basic knowledge to distinguish between types of bot accounts and non-bot accounts is required.

Conventionally, identifying bot accounts and not bots can be carried out by observing the activity pattern on an account. For example, noticing that a particular account does more retweeting than creating original tweets, writes many tweets, but only has a few followers. Besides, the account also does not have a biography, a profile picture, and writes the same tweet content as another user at the same time. However, such cognitive approaches are rated inefficient and merely focus on

precision.

Therefore, an approach to detecting bots with machine

learning is created. The machine learning models are used because of their capacity to analyze extensive data based on parameters. According to a study, the random forest algorithm has the best accuracy, which is 95%, compared to Naïve Bayes Multinomial algorithm (70%), Naïve Bayes Gaussian algorithm (68%), and Logistic Regression algorithm (52%) .

Therefore, the random forest classifier algorithm is selected in this paper. The system classified the multiclass classification by classifying accounts into four different classes, which are human, informative, spammers, and fake followers. Then, a completed machine learning model was inserted into the web application. The website serves as an interface for end-users to use machine learning systems. In the end, the system underwent a deployment process to the cloud service.

**Conclusion :-** We came to know the introductory part of Twitter bots and its specifications . Thus its is the need of an hour to detect and avoid bots on Twitter to be safe from malicious activities .

**Assignment No. 2**

**Title of Assignment :-**  State the requirements and specifications of the project .

**Theory :-**

A. Classification Type

Classification is a process of assigning a category or label

that has been defined as data that does not yet have a category.

In general, there are three types of the data classification process, namely binary, multiclass, and multilabel classification .

•Binary classification is a process of classifying each

Element in a group into two groups or categories.

•Multiclass classification is a classification process

involving more than two classes. However, the multiclass

classification creates an assumption that each given sample is categorized into a single label (mutually exclusive).

B. Machine Learning

Machine learning is a technique that enables the system to

learn from data compared to using direct programming so that it can deliver relevant results .

C. Random Forest Algorithm

Random forest was first introduced by Leo Breiman . The random forest classifier is the development of the decision tree. It consists of a combination of many decision trees, with each tree relying on independent random vector values with an equivalent distribution of each tree .

D. Twitter Social Media

Twitter is a micro-blogging social network that allows its

users to send and read short messages up to 140 words, which are then called tweets . Jack Dorsey founded this social media in 2006. Unlike social media such as Facebook or MySpace, on Twitter, the relationship between to follow an account and the followers are not reciprocal. It means that an account can follow other accounts without automatically be followed by the account it follows.

E. Bots and Twitter Bot Types

In general, bot means an application that performs tasks automatically. In social media, bot domain is a social media accounts programmed to perform social media activities automatically, so they look like real humans. According to research from the University of Southern California, at least 9% to 15% of active Twitter users are bots . Until 2017, there were 319 million active users each month. It means there are almost 48 million bot accounts spread on the Twitter social network.

Factors that influence bot growth include Twitter API

support, bot development cycles that can be created quickly,

Twitter public platforms, and the flexibility to create as many accounts as possible. According to the Digital Forensic Research (DFR) of the Atlantic Council Lab, there are several features indicating that an account is a bot, including amplification, anonymity, activity, similarity, and description of "bot" in the account . Whereas the Twitter bot types based on account activity are as follows.

• Informative, i.e., a bot that functions to disseminate

information to users. For example, bots that publish facts,

earthquake information, and write poetry content as well as humor content .

• Spammers, i.e., bots that work to broadcast spam conten

• Fake Followers, i.e., bots that act as shadow followers for an account. The purpose of using fake followers is tocreate an image that an account seems to have prominent

**Conclusion :-** We came to know about the different requirements and specifications needed to build a bot detection system .

**Assignment No. 3**

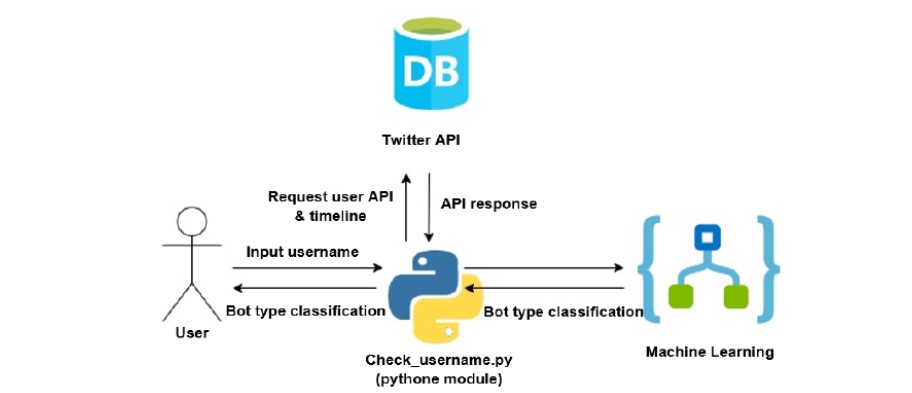
**Title of Assignment :-** Draw the Flowchart / block diagram of your mini project .

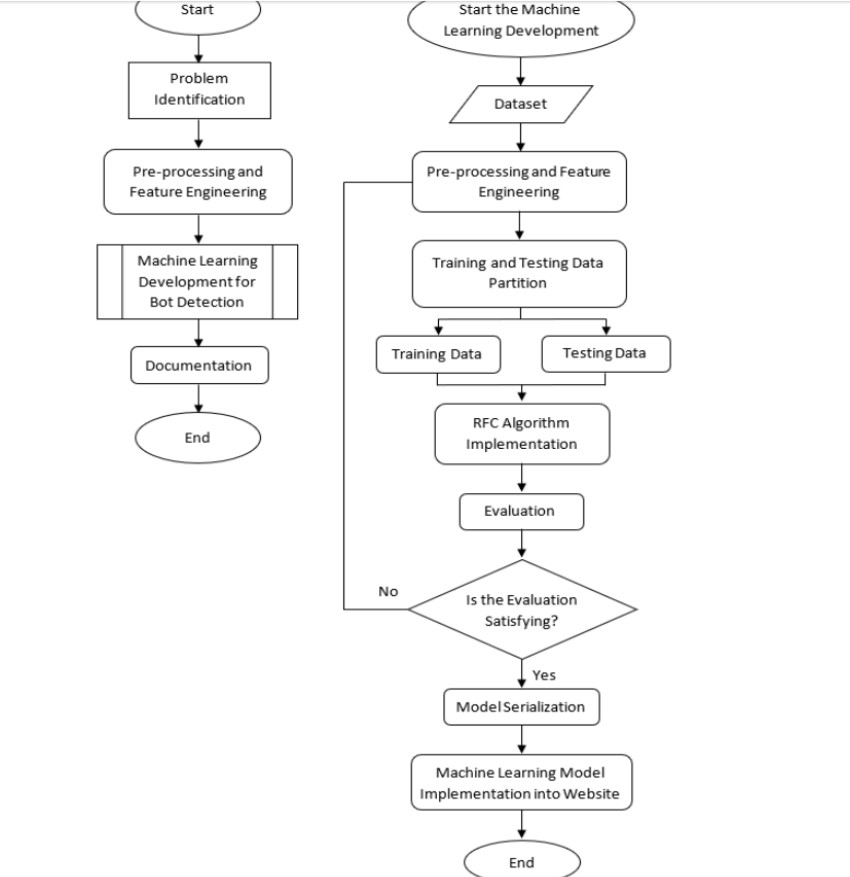
**Theory :-** The initial process in the system development was by collecting data. The labeled username data were downloaded to a public repository.

Then, a crawling process was carried out to obtain additional data such as metadata, tweets, and other attributes using the Tweepy library.

The data collection process employed the Jupyter Notebook as

an Integrated Development Environment (IDE).





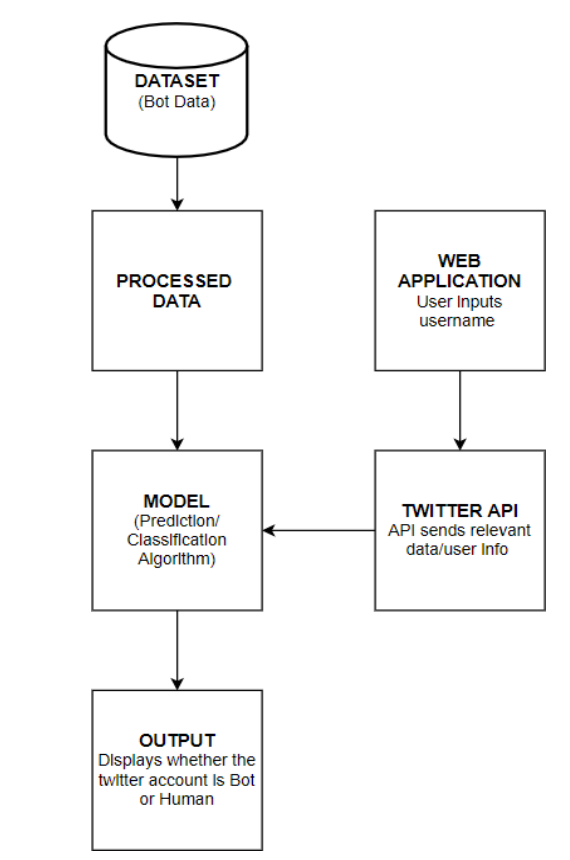
**Conclusion :-** We came to know about the Block diagram and the flowchart of our mini project . The set of flow of activities was cleared through it .

**Assignment No. 4**

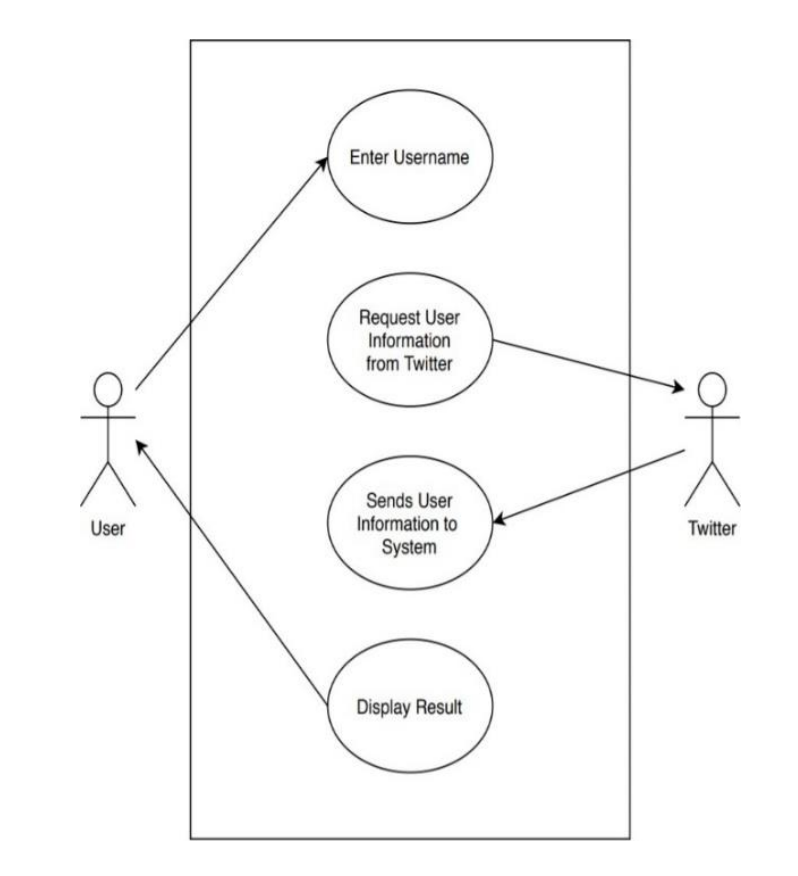
**Title of Assignment :- System Architecture.**

**Theory:**

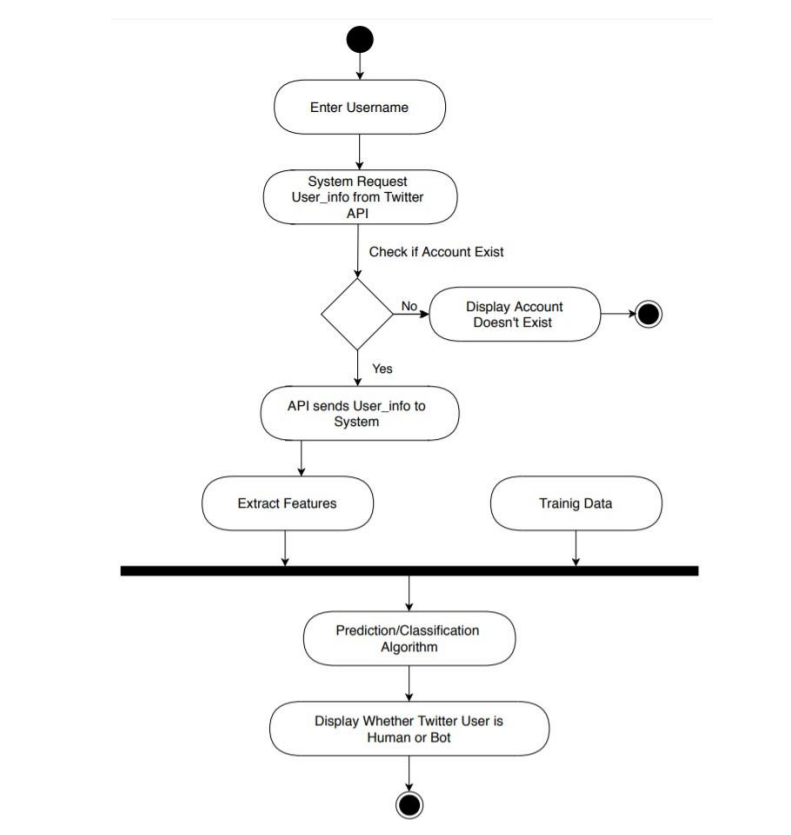
**1) Block Diagram:** A block diagram is a specialized, high- level flowchart used in engineering. It is used to design new systems or to describe and improve existing ones. Its structure provides a high-level overview of major system components, key process participants, and important working relationships.



**2) Use case Diagram:** Use case diagrams are used to gather the requirements of a system including internal and external influences. These requirements are mostly design requirements. Hence, when a system is analyzed to gather its functionalities, use cases are prepared and actors are identified. When the initial task is complete, use case diagrams are modelled to present the outside view.



**3) Activity Diagram:** Activity Diagrams describe how activities are coordinated to provide a service which can be at different levels of abstraction.



**Conclusion:** We came to know about DFD, Use case & Activity diagram of our mini project.

**Assignment No. 5**

Title of Assignment: Model

**Theory:** To build our model, we first trained our dataset on different algorithms to find the most optimum algorithm to go forward with. The algorithms used are:

1. Decision Tree
2. Random Forest

Before passing the dataset to these algorithms, the dataset was pre-processed. Then the most efficient algorithm was selected to be the model for our system. We found the most efficient and accurate algorithm was the random forest algorithm. Random forest is a collection of decision trees.

It generally provides higher accuracy than the normal decision tree as was observable in our experiment too. It used different attributes like friends count, followers count, listed count, etc. to predict the result as a bot or a genuine account. Eighty percentage of the dataset was used to train the model and the rest twenty percentage to test the trained model.

Most of the time, the model predicted the account as a bot whenever the verified attribute was ‘FALSE’. It’s not necessary for accounts specified as ‘FALSE verified’ to be a bot and vice versa. Also, most of the bot accounts predicted were not popular. That’s because these bot accounts generally remain in a stealth mode.

The model built using the random forest algorithm was quite efficiently able to predict the account as a bot or a genuine account

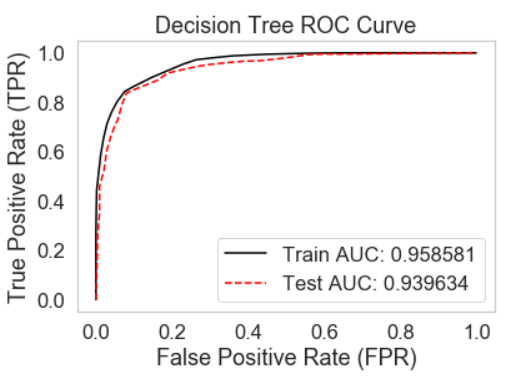
**Conclusion:** We came to know about efficient algorithms for our mini project.

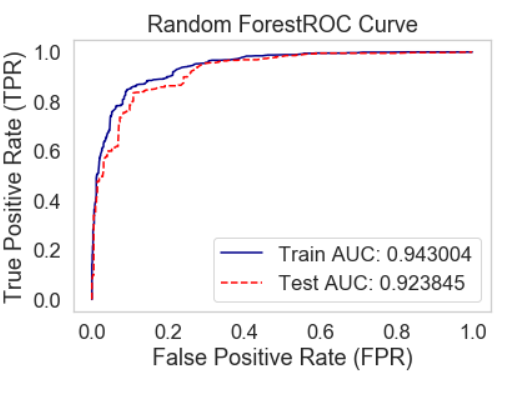
**Assignment No. 6**

**Title of Assignment: Result**

**Theory:** The model required for this project aims to predict whether a given Twitter account is a genuine user or a bot and also to have the best accuracy. For this reason, we implemented various types of classification algorithms and selected the one with the maximum accuracy score. The below table shows the algorithms and their accuracy scores.

|  |  |  |
| --- | --- | --- |
| **Algorithms** | **Accuracy Rate** | **True Positive Rate** |
| Decision Tree | 87.85% | 86.03% |
| Random Forest | 85.82% | 83.50% |





As it is clear from the table that random forest has the highest score hence we used random forest in our model.

Compared against bots our model is 85.82% accurate, with a misclassification rate of 14.72% The true positive rate of this model for bots is 83.50%. This was done with a total dataset of 2,748 total accounts; 1,361 bots and 1,387 real accounts.

**Conclusion:** We came know about the accuracy of our project.