**Twitter Spam Detection**

A Project Report submitted in partial fulfillment of the requirements for the award of the degree of

**Bachelor of Technology**

### in

**Computer Science and Engineering/ Electronics and Communication Engineering**

by

### Group Members’ Names:

### Ajay Walke (112015006)

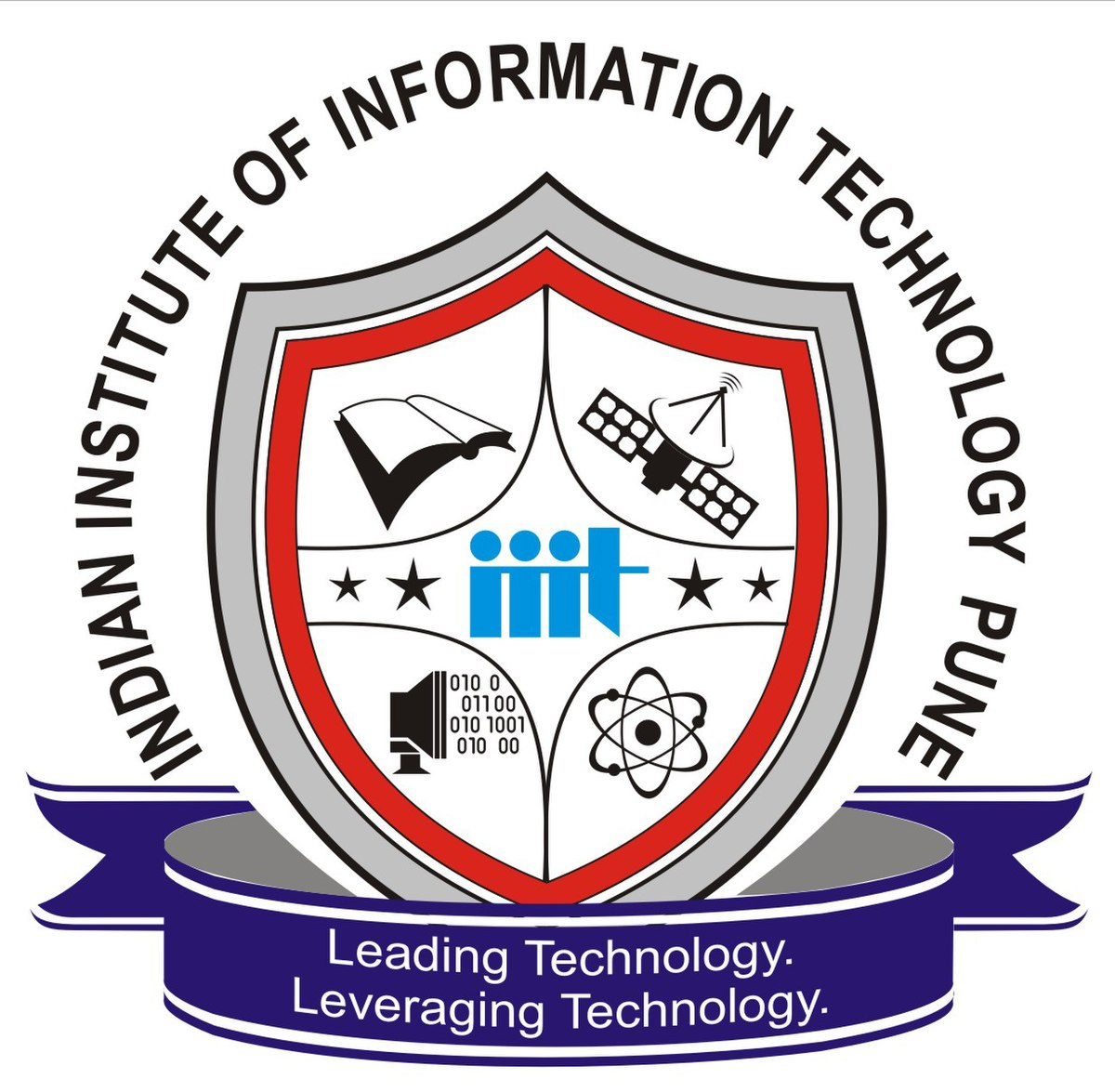
### Rohan Khavale (112016031)

### Manas Agarwal (112015005)

### Sahil Thakare (112015153)

**Under the Supervision of: Prof. Ritu Tiwari**

### Semester: Fifth

****

#### Name of Department: Department of Computer Science and Engineering/ Department of Electronics and Communication Engineering

#### 

#### Indian Institute of Information and Technology, Pune

**(An Institute of National Importance by an Act of Parliament)**

#### DECEMBER 2022

**BONAFIDE CERTIFICATE**

This is to certify that the project report entitled **“Twitter Spam Detection”** submitted by **Ajay Walke** bearing the **MIS No: 112015006**, **Rohan Khavale** bearing the **MIS No: 112016031**, **Manas Agarwal** bearing the **MIS No: 112015005**, **Sahil Thakare** bearing the **MIS No: 112015153**, in completion of his/her project work under the guidance of **Prof. Ritu Tiwari** is accepted for the project report submission in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in the **Department of Computer Science and Engineering/ Electronics and Communication Engineering**, Indian Institute of Information Technology, Pune (IIIT Pune), during the academic year **2022-23**.

#### Supervisor’s Name HoD Name

Prof. Ritu Tiwari Dr. Tanmoy Hazra

Dean Assistant Professor

Department of CSE/ECE Department of CSE/ECE

IIIT Pune IIIT Pune

Project Viva-voce held on

**Undertaking for Plagiarism**

We **Ajay Walke, Rohan Khavale, Manas Agarwal, Sahil Thakare** solemnly declare that research work presented in the **report/dissertation** titled **“Twitter Spam Detection using the machine learning algorithms”** is solely **our** research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete report/dissertation has been written by **us/me**. I understand the zero-tolerance policy of **Indian Institute of Information Technology Pune** towards plagiarism. Therefore, **we** declare that no portion of my **report/dissertation** has been plagiarized and any material used as reference is properly referred/cited. I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of the degree, the Institute reserves the rights to withdraw/revoke my **B. Tech** degree.

**Student’s Name and Signature with Date**

**Ajay Walke**

**Rohan Khavale**

**Manas Agarwal**

**Sahil Thakare**

**Conflict of Interest**

**Manuscript title: Social Network Spam Detection using Machine learning**

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers’ bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

**Student’s Name and Signature with Date**

**Ajay Walke**

**Rohan Khavale**

**Manas Agarwal**

**Sahil Thakare**

Problem statement and objectives

**Problem Statement:**

Detection of social network spam (In Twitter) using Machine Learning.

**Problem Objectives:**

To detect the spammers in twitter interface using the numerical features with the help the various machine learning algorithms. To prevent unwanted, malicious, unsolicited content or behaviour, manifested in various ways that are caused by the spammers including microblog, message, malicious links, fraudulent reviews etc. To analyze the effectiveness of the research, the work has been computed with existing work and it has been computed with existing work and it has been concluded that the value of precision, recall and F-score of the research has been increased.

## ACKNOWLEDGEMENT

This project would not have been possible without the help and cooperation of many. I would like to thank the people who helped me directly and indirectly in the completion of this project work.

First and foremost, I would like to express my gratitude to our honorable Director, **Prof. O.G. Kakde**, for providing his kind support in various aspects. I would like to express my gratitude to my project guide **Prof. Ritu Tiwari**, **Department of CSE/ECE**, for providing excellent guidance, encouragement, inspiration, constant and timely support throughout this **B. Tech Project**. I would like to express my gratitude to the **Head of Department (Dr. Tanmoy Hazra)**, **Department of CSE**, for providing his kind support in various aspects. I would also like to thank all the faculty members in the **Department of CSE/ECE** and my classmates for their steadfast and strong support and engagement with this project.

Abstract

## Abstract

Twitter spam has long been a critical but diﬃcult problem. to be addressed. So far, researchers have developed a series of machine learning-based methods and blacklisting techniques to detect spamming activities on Twitter. Previous research has suggested a few detection and defense methods that secure Twitter users from spammers. So, we are going to work on spam detection techniques on Twitter. This study consists of 3 sections: 1- Background about spam detection on Twitter. 2- A literature review comparative analysis of machine learning, deep learning and hybrid algorithms. 3- Discussion on limitations of previous studies and future directions. According to our investigation, current methods and techniques have achieved accuracy of around 80%. Using the API methods provided by Twitter, we crawled active Twitter users, their followers/following information and their most recent 100 features. Then, we evaluated our detection scheme based on the suggested user and content-based features.

**Keywords:** Twitter, Spammer, Deep Learning, Machine learning algorithm

## TABLE OF CONTENTS

**Abstract 6**

**(i)** [**List of Figures/Symbols/Nomenclature 8**](#_heading=h.gjdgxs)

**(ii)** [**List of Tables**](#_heading=h.30j0zll) **9**

1. **Introduction 10**

[1.1 Overview of work . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1](#_heading=h.1fob9te)0

[1.2 Motivation of work . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1](#_heading=h.3znysh7)1

[1.3 Literature Review. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1](#_heading=h.3znysh7)2

1.4 Research Gap. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

1. **Problem Statement**

2[.1 Research Objectives. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1](#_heading=h.3znysh7)4

2[.2 Methodology of work. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1](#_heading=h.3znysh7)5

1. **Analysis And Design 19**
2. **Results and Discussion 20**
3. **Conclusion and Future Scope 22**

**6 References 23**

List of Figures/ Symbols/ Nomenclature

**List of Figures/Symbols/Nomenclature**

**List of Figures:**

1. Activity Diagram. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15
2. Data Flow Diagram. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15
3. Framework of model. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16
4. Features in the Dataset. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16
5. Designing of the spammer detection. . . . . . . . . . . . . . . . . . . . . . . 19
6. Performance of Naïve Bayes. . . . . . . . . . . . . . . . . . . . . . . . . . . . 20
7. Performance of KNN. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20
8. Comparative Study of all algorithms. . . . . . . . . . . . . . . . . . . . . . . 21
9. Classification of user-id as spammer and non-spammer. . . . . . . . . . . 21

List of Tables

## List of Tables

1. Literature Survey Table . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

**Chapter 1**

**Introduction**

## Overview of Work:

# In the following report we discuss how our team has worked on Spam Detection on Social Media Network Twitter. Our goal is to prevent unwanted, malicious, unsolicited content or behavior, manifested in various ways including microblogs, messages, malicious links, fraudulent reviews, etc. to analyze the effectiveness of the research, the work has been computed with existing work and it has been concluded that the value of precision, recall and F-measure of the research work has been increased. We used various algorithm like KNN,SVM, Naïve Bayes

Literature Review

## Motivation of the Work:

## During lockdown we all were continuing our life in online mode whether it be work or talking to our dear ones we all we dependent on Internet but also, we all were coming across many spam messages which were misleading and spreading rumors about the ongoing pandemic which made it more difficult to handle the situation. So, we as a team decided to target this issue and control the spammers. To do this we did a lot of intensive literature surveys and selected twitter as the main social networking platform to test our spam detection model. And now after months of hard work we can achieve 96-99% accuracy in spam detection and it gives our team great honor to get working in the real world.

Literature Review

## Literature Review

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of research paper** | **Author & Year** | **Methodology** | **Results** |
| Twitter Spam detection based on deep learning | Tingmin Wu, Shigang Liu  (January 2017) | -Deep Learning | - The evaluation of this paper is mainly on empirical studies. We will carry out theoretical studies on the outperformance of our methods in order to better understand the deep-learning based spam detection framework. |
| Real-Time Twitter Spam Detection and Sentiment Analysis using Machine Learning and Deep Learning Techniques | Anisha P Rodrigues, Roshan Fernandes, Aakash A, Abhishek B, Adarsh Shetty, Atul K, Kuruva Lakshmanna  (2017) | -LSTM  -CNN  -RNN | - The classification results showed that the features extracted from the tweets can be satisfactorily used to identify if a certain tweet is spam or not and create a learning model that will associate tweets with a particular sentiment. |
| Twitter Spammer Detection | Ashwini Bhangare, Smita Ghodke, Kamini Walunj, Utkarsha Yewale  (2018) | -SVM  -Nature Language Processing | - NLP has implemented with handling dataset in well manner. |
| Detection of Social Network Spam Based on Improved Extreme Learning Machine | ZHIJIE ZHANG, RUI HOU, AND JIN YANG  (2020) | - I2FELM | -To address insufficient labeled data in the social network, semi-supervised learning method will be substituted into I2FELM model to detect Twitter spam automatically based on a small amount of labeled data. |
| Twitter Spam Detection Using Machine learning | Asfa Falak, Dr. Hamid Ghous, Dr. Mubashir Malik  (2021) | -SVM  -KNN | -Conduct more theoretical studies on the out performance of methods better understand the social honey pots based  on malicious user’s detection framework. |
| Spam detection on Twitter using a support vector machine and users’ features  by identifying their interactions | Saleh Beyt Sheikh Ahmed, Mahnaz Rafie, Seyed Mojtaba Ghorabie  (2021) | -Gaussian  -polynomial kernels  -MLP  -Naïve Bayes | -The proposed method is compared with multi-layer perceptron (MLP), Naive Bayes (NB), random forest (RF), and k-nearest neighbors (KNN) methods in terms of standard criteria. |

Research Gap

**1.4 Research Gap**

#### First, other factors will be considered to accurately identify spam (e.g., semantic analysis and emotion analysis). We also plan to use the feature selection and oversampling method to select appropriate feature sets and improve model fitting. On the other hand, the current model will be replaced by a semi-supervised learning method to automatically detect Twitter spam based on a small amount of labeled data.

#### In previous work, more variables needed to be added in the framework to enhance the accuracy of the model and classification rate. Need to improve text similarity for extracted new strange words from tweets. In previous research[3], data mining algorithms were applied on small amounts of collected dataset and limited tweets. So, large amounts of data sets need to be tested for the accuracy of previous algorithms. In Future, we can collect the dataset of tweets in different languages. We can apply data mining algorithms on other social media platforms like Facebook, Instagram, LinkedIn, YouTube, and WhatsApp. More classifiers can be added that can make Twitter spam detection more valuable for users. Research will help to solve model scalability without performing comparative accuracy. Can use the characteristics of spammers at different levels of granularity have been used by some interesting patterns released by spammers.

#### The performances for all four metrics on for datasets are better than other all the time. As shown in Figure 6[4], the F-measure is much higher than others, with averagely 30% higher than Random Forest and almost nine times of Naive Bayes in Dataset 2 and 4. Even the Decision Tree method achieves almost the same as our method at Dataset 1, it only remains half when testing on Dataset 4[4].

# 

# Chapter 2

# Problem Statement

## Social Network Spam detection

Detection of social network spam (In Twitter) using Machine Learning. Use of the various machine learning algorithms to find the spammers in the Twitter and classify the user based on the numerical features available in the dataset.

1. **Research Objectives**

# To judge authenticity of a user.

# To prevent unwanted, malicious, unsolicited content or behavior, manifested in various ways including microblogs, messages, malicious links, fraudulent reviews, etc.

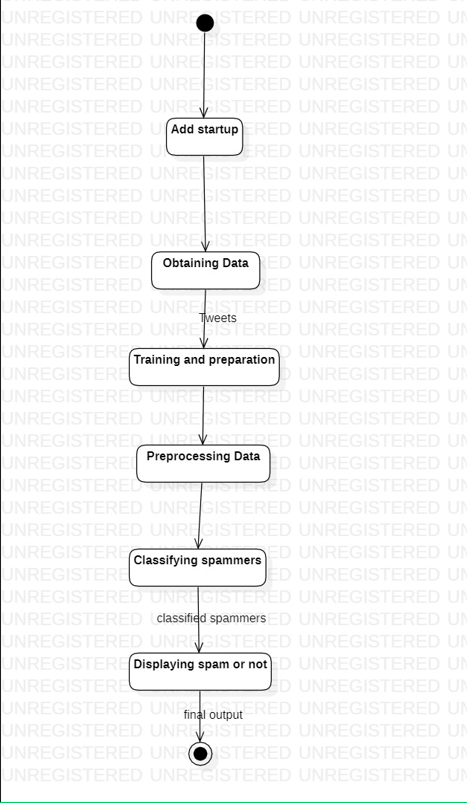
# To analyze the effectiveness of the research, the work has been computed with existing work and it has been concluded that the value of precision, recall and F-measure of the research work has been increased.

# To detect the spammers in twitter interface using the numerical features with the help the various machine learning algorithms.

Methodology

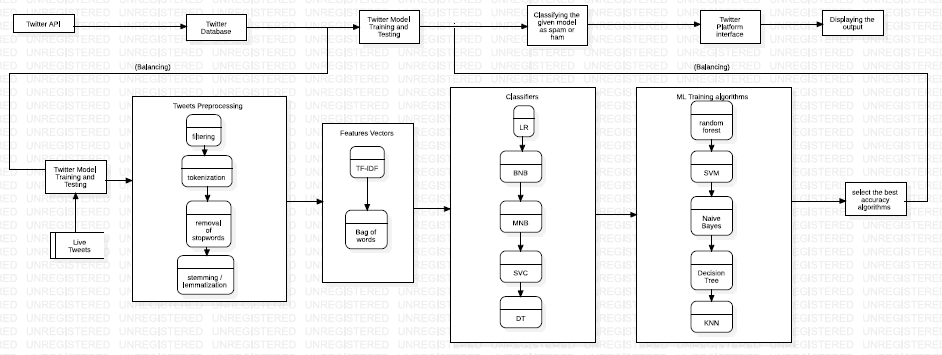
1. **Methodology of the Work**

The workflow of the Twitter spam detection model is as follows:



**Fig. 1 Activity diagram**

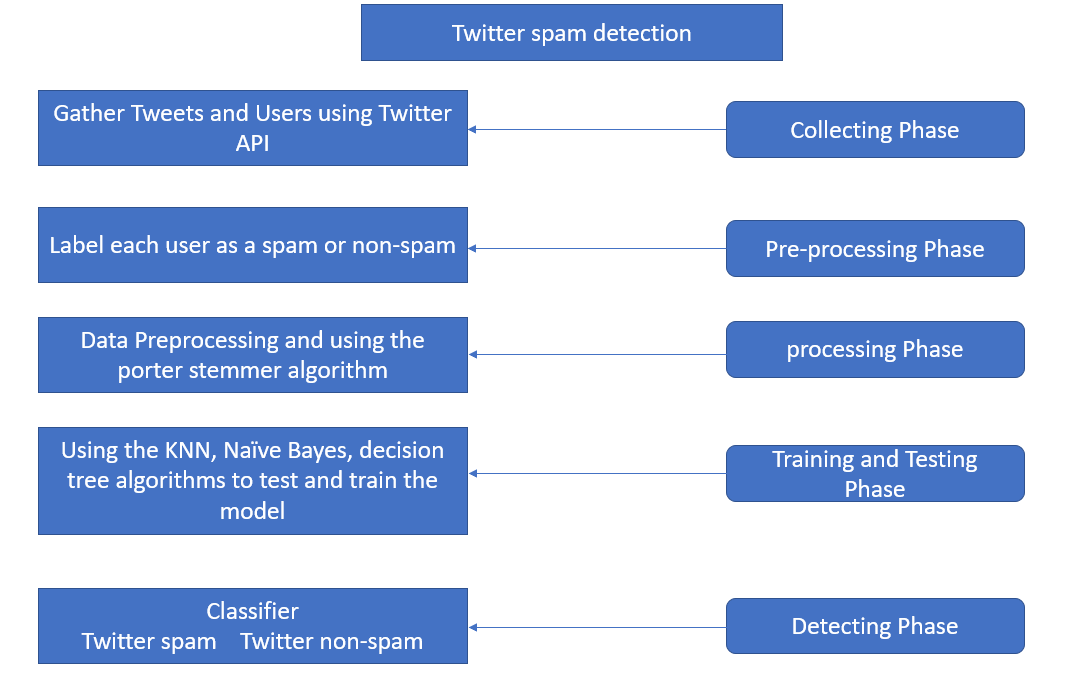
It consists of the database for the training and testing of the model. Various classifiers are used for the internal computations. The dataflow in the whole document is as follows:

****

**Fig. 2 Data Flow diagram**

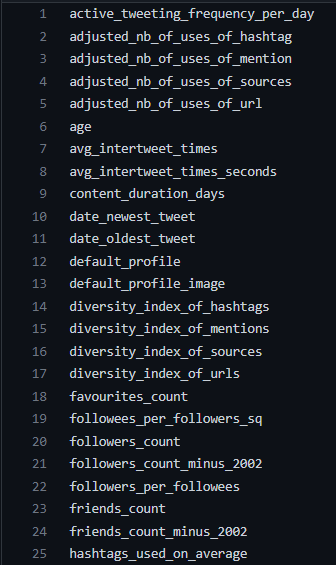
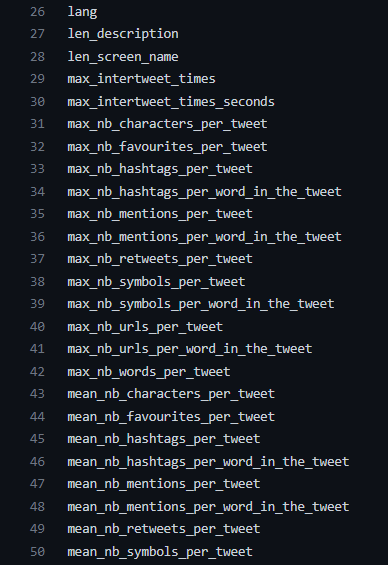
The model is divided into phases to find the spammer in twitter. Each phase has its one importance, and the improvement of the model is done at each phase.

Methodology

****

**Fig. 3 Framework of model**

The dataset consists of the various kind of the features including the number of followers, active frequency on twitter per day, total number of the post etc. Following figures consists of the features used in the dataset to train the model.

**Fig. 4 Features in the Dataset**

The model is trained based on the dataset gathered from twitter using the third-party app. Twitter does not provide direct access to the dataset the gathered is done by the tweepy library in python. The whole project consists of the core concept like the data preprocessing of the dataset, standardization, normalization, vectorization, tokenization, and stemmer algorithm for the processing of the text of dataset. The centralized concept of the model includes the use of various algorithms and the various optimization techniques to train the model. Following methodology is used in the model:

Methodology

1. Data Processing: Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. While creating this model, it is not always a case that we come across clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put it in a formatted way. So, for this, we use data preprocessing tasks. Different kinds of data preprocessing techniques are used on the dataset.
2. Splitting the dataset: The dataset is divided into the training and testing set. The first dataset i.e, user-id is divided into training dataset user-id and testing user-id. The second dataset i.e, features is divided into training and testing set containing all features set.
3. Removing the irrelevant data: The training and testing set of both the dataset contains the irrelevant and null values that are creating a huge mess while developing the dataset. The features containing the null values are filtered using the porter stemmer algorithm and the dataset is manipulated accordingly. This filtered dataset is then used for the next processes.
4. Standardization of the data: Data standardization is the process of rescaling the attributes so that they have mean as 0 and variance as 1. The goal to perform standardization is to bring down all the features to a common scale without distorting the differences in the range of the values. Standardization is used on the training dataset before applying any algorithm.
5. Naïve bayes: Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts based on the probability of an object.
6. KNN: K-Nearest Neighbors is one of the simplest Machine Learning algorithms based on Supervised Learning technique. The K-NN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most like the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.
7. SVM: Support vector machine (SVM) proposed by vapnik and Cortes have been successfully applied for gender classification problems by many researchers. An SVM classifier is a linear classifier where the separating of the hyper plane is chosen to minimize the expected classification error of the unseen test patterns. SVM is a strong classifier which can identify two classes. SVM classifies the test image to the class which has the maximum distance to the closest point in the training. SVM training algorithm built a model that predicts whether the test image falls into this class or another. SVM requires a huge amount of training data to select an effective decision boundary and computational cost is very high

Methodology

even if we restrict ourselves to single pose (frontal) detection. SVM is a learning algorithm for classification. It tries to find the optimal separating of the hyperplane such. The SVM training algorithm built a model that predicts whether the test image falls into this class or another.

1. Decision Tree: Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules, and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed based on features of the given dataset.
2. Random Forest: Random Forest is a classifier that contains a few decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.
3. Multilayer Perceptron: Multi-layer perception is also known as MLP. It is fully connected dense layers, which transform any input dimension to the desired dimension. A multi-layer perception is a neural network that has multiple layers. To create a neural network, we combine neurons together so that the outputs of some neurons are inputs of other neurons. A multilayer perceptron (MLP) is a feed forward artificial neural network that generates a set of outputs from a set of inputs. An MLP is characterized by several layers of input nodes connected as a directed graph between the input nodes connected as a directed graph between the input and output layers. MLP uses backpropagation for training the network. MLP is a deep learning method.
4. Grid Search CV: Grid-searching is the process of scanning the data to configure optimal parameters for a given model. Depending on the type of model utilized, certain parameters are necessary. Grid-searching does NOT only apply to one model type. Grid-searching can be applied across machine learning to calculate the best parameters to use for any given model. It is important to note that Grid-searching can be extremely computationally expensive and may take your machine quite a long time to run.

# Chapter 3

# Analysis and Design

# The dataset is divided into two parts, one consists of the user-id of the user collected from the twitter API. These first data sets is divided into two training and testing sets. Another dataset consists of features with numerical values. These are unprocessed datasets consisting of various irrelevant data.

# The whole data set is analyzed on the basis on the relevancy of the features and their effect on the improvement on the model. All the columns consisting of the null values is removed and the standardization and normalization is used to improve the dataset. Total is algorithm are used for designing the algorithm naïve bayes, KNN, SVM, Decision tree, Random Forest algorithm, multilayer perceptron algorithm to match the research gap. Framework of the model is based on the improvement the initial applied model by using the successor algorithm. Finally whole structure is presented in format of the graph to visualized the results and then see the total significant on the identification of the spammer of the twitter. Whole model structure is divided into following part and flow is as follows:

# 

# 

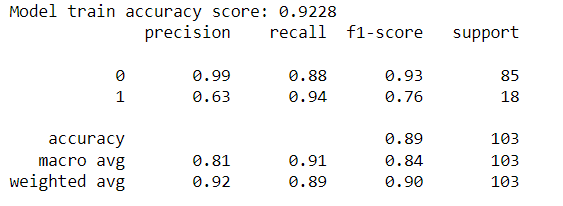
**Fig. 5 Designing of the spammer detection**

# Chapter 4

# Results and Discussion

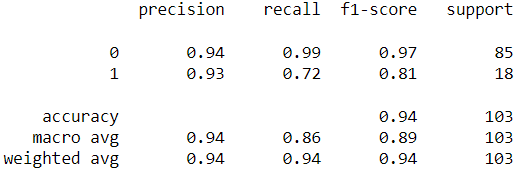
All the algorithms have given the best improvement to the model in the initial phase of the algorithms model was trained by the naïve bayes, and lastly the multilayer perceptron is used the improve the model the model. Following the details analysis of the all 6 algorithms that are used in the model and the detail studied of the algorithms and their impacts are also mentioned.

1. Naïve bayes: Naïve bayes implemented on the initial phase of the algorithm. Initial phase was quite time consuming and the accuracy of the model was 92%. Also, the performance metrics was quite satisfying.



**Fig. 6 Performance of Naïve Bayes**

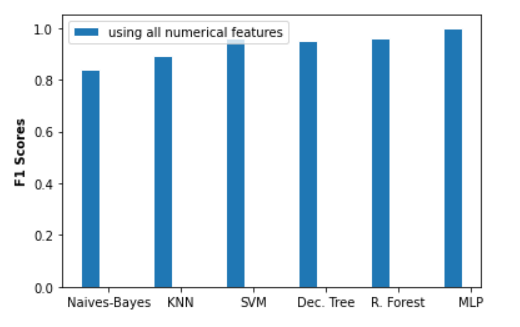
1. KNN: KNN is used for the dimensionality reduction as the dataset consisting of the user id, features with the numerical values are of large weights consisting of the irrelevant features. After implementing the KNN the model accuracy improved to the 93%.



**Fig. 7 Performance of KNN**

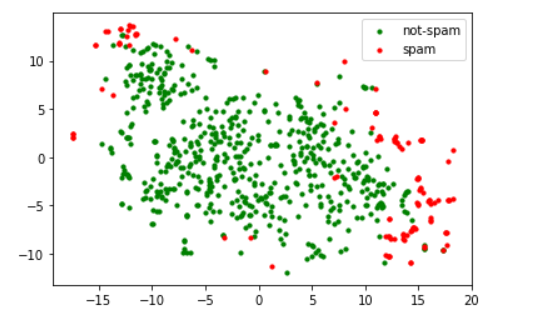
1. SVM: SVM has boosted the score with the relevant changes in the model with accuracy 94%.
2. Decision Tree: Decision tree also increased the accuracy with the significant figures.
3. Random Forest algorithm: Random Forest algorithm also measuredly improved performance to 95%.
4. Multilayer-perceptron model: MLP is used to optimal improved the accuracy by boosting the weak learners and final accuracy is nearly about 96%.
5. Grid search CV: It is used to choose the optimal hyperparameters to trained the model and the changed the model accordingly. It is used to measures accuracy in each algorithm

Following is the comparative study on the accuracy of the various algorithms:



**Fig. 8 Comparative Study of all algorithms**

Finally, the dataset consisting of the all the user-id are classified as the spammer or not spammer account



**Fig. 9 Classification of user-id as spammer and non-spammer**

# Chapter 5

# Conclusion and Future Scope

This study presents a novel Twitter spam detection method and new insights into the sophisticatedly evolving techniques for spamming on Twitter. In which the features set consists of user attributes, content, activity, and relationships in online social networks for identifying the real spam. It was also shown that automated spam accounts follow a well-defined pattern with surges of intermittent activities. Spammer Detection has a strong commercial interest because companies or individuals wants to improve the security on social media. During the analysis of the data, we observed that spam users tend to be selective in following other users thereby forming enclaves of spammers. This is a high-level observation that we aim to explore further in the future. Additionally, both the two broad user groups, i.e, human users, and social bot (autonomous entity) users contain spammers, whose spamming behaviour tends to be similar. The distinction between legitimate human users vs. legitimate social bots as well as human spammers vs. social bot spammers needs to be investigated further. Another interesting dimension for future work is to study the effect of the recent increase in the maximum length of tweets on spamming activity. Intuitively, automated spam accounts will face difficulties in generating lengthier tweets intelligently, thereby making these tweets easier to identify. Though the model accuracy depends on the features that are taken from twitter. Some real-life scenarios improve the spamming techniques according the features. Model is based on the sequential implementation of the algorithms, error in the any phase causes the failures in the model.

# References

1. Tingmin Wu, Shigang Liu, “Twitter spam detection based on deep learning,” in IEEE Access, 2017
2. Anisha P Rodrigues, Roshan Fernandes, Aakash A, Abhishek B, Adarsh Shetty, Atul K, Kuruva Lakshmanna, “Real-Time Twitter Spam Detection and Sentiment Analysis using Machine Learning and Deep Learning Techniques,” in IRJET Access, 2017
3. Ashwini Bhangare, Smita Ghodke, Kamini Walunj, Utkarsha Yewale, “Twitter Spammer Detection,” IRJET Access, 2018
4. Z. Zhang, R. Hou and J. Yang, "Detection of Social Network Spam Based on Improved Extreme Learning Machine," in IEEE Access, vol. 8, pp. 112003-112014, 2020, doi: 10.1109/ACCESS.2020.3002940.
5. Asfa Falak, Dr. Hamid Ghous, Dr. Mubashir Malik, “Twitter Spam Detection Using Machine learning,” IJSER, 2021
6. Saleh Beyt Sheikh Ahmad & Mahnaz Rafie & Seyed Mojtaba Ghorabie, "Spam detection on Twitter using a support vector machine and users’ features by identifying their interactions", in IEEE Access, 06 January 2021.
7. Rumi Juwairiyyah, Nanditha Sriram, Jyotshna Bhushan Sharma, Babeethabeetha, “Social media spam detection using deep learning,” IRJET Access, June 2020.
8. Deepali Prakash Sonawane Dr. Baisa L. Gunjal, “New Approach for Detecting Spammers on Twitter using Machine Learning Framework,” in IEEE Access, May 31st 2020 2020.
9. M Mostafa, An Abdelwahab, H M Sayed, “Detecting spam campaign in twitter with semantic similarity,” in IEEE Access, 2020.
10. Francesco Pierri, Carlo Piccardi, Stefano Ceri, “A multi-layer approach to disinformation detection on Twitter,” IRJET Access, 2020.