**CLASSIFICATION OF VIRTUAL HARASSMENT ON SOCIAL NETWORKS USING ENSEMBLE LEARNING**

**By**

**ISIEKWENE CHINYERE CHIOMA**

**209074070**

**In partial fulfilment of the award of Masters of Science in Computer Science,**

**Submitted to the Computer ScienceS Department, University of Lagos, Akoka, YABA, LAGOS, NIGERIA.**

**FEBRUARY, 2023**

**Certification**

This is to certify that this research work was carried out by **ISIEKWENE CHINYERE CHIOMA** with **Matric no: 209074070** a student of the department of Computer Science, University of Lagos, Akoko, Yaba, and same has been duly supervised and approved in partial fulfillment of the requirements for the award of Masters of Science in Computer Science, University of Lagos State, Akoka, Yaba.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Dr. N. A. Azeez Date**

Project *Supervisor*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**PROF. A. P. ADEWOLE Date**

Head of Department

**Dedication**

The work is dedicated to the God Almighty, the Giver of life and the Maker of everything, the Anchor of my life, the Author and Finisher of our Faith, I am who I am, the Pillar that holds my life, the Refiner, the Builder, and the Sustainer, for His grace upon me. He is YAHWEH.

**Acknowledgement**

My heartfelt gratitude goes to GOD Almighty for the gift of life, divine guidance, understanding, protection, and provision throughout the program. I am indebted to my intelligent, articulate, resourceful, cheerful, understanding, approachable, vast, humble, and honorable supervisor Dr. N. A. Azeez. To my ever-supportive husband and sponsor, John O. Isiekwene, words cannot express how I feel right now, thank so much for all you do, God bless you greatly. And to the best children (Baruch, Benaiah, and Bukkiah) and siblings (Ezinne, Ugonma, Abraham, Oluchi, Moses, and Deborah) ever, I love you all. To my dearest Mother, Thank you for your continuous guidance and tutelage. God bless you Ma. Dear Daddy, six years now since you left us and it still feels like yesterday. I know you are smiling down at your ever-resilient, resources, multi-talented Chi-chi. This is for you Dear Ikechi C. Eziuku aka Gadaffi, continue to rest in peace (Amen). To my mentors and friends Dr. U. C. Ogude, Mr. Oludipe and Mr. Donaldson (YCT), Sadat, Tomi, Modupe, Juan, Pamela, Odunola, Damilola, Remi, my colleagues, etc. God bless you all for me.

**ISIEKWENE CHINYERE CHIOMA (NEE EZIUKU)**

**209074070**

**TABLE OF CONTENTS**

Title page i

Certification ii

Dedication iii

Acknowledgement iv

# **CHAPTER ONE**

# **INTRODUCTION**

## 1.0 Background to the study: 1

## 1.1 Statement of the Problem 4

## 1.2 Aim and Objectives 5

1.2.1 Aim 5

1.2.2 Objectives 5

## 13 Contribution to Knowledge 5

## 1.4 Limitations of the Study 5

# **CHAPTER TWO**

# **LITERATURE REVIEW**

## 2.1 Virtual Harassment 6

## 2.2 History of Virtual Harassment 6

### 2.2.1 Harassment - definition 6

### 2.2.2 Origins of Harassment 7

## 2.3 Traditional/Old-fashioned Harassment verses virtual harassment 7

### 2.3.1 Traditional/Old Fashioned Harassment 7

2.3.2 The features of virtual harassment include: 8

2.4 A Technological Evolution 9

2.5 Related Works 10

**CHAPTER THREE**

**METHODOLOGY**

3.0 Data Collection 20

3.0.2 Dataset 2 20

3.0.3 Dataset 3 21

## 3.0.4 Dataset 4 21

## 3.1 Data Preprocessing 22

## 3.2 Featured Engineering 22

## 3.1.1 Activity Features 23

## 3.1.2 User Feature 23

## 3.1.3 Content Features 24

## 3.3 Machine Learning Algorithms 24

## 3.4 Ensemble Learning 31

## 3.5 Experiments 38

## 3.6 Performance Metrics 38

# **CHAPTER FOUR**

# **SYSTEM IMPLEMENTATION AND TESTING**

## Data Analysis for Dataset 1 43

## 4.1 Data Analysis for Dataset 2 47

## 4.2 Data Analysis for Dataset 3 48

## 4.3 Data Analysis for Dataset 4 55

## 4.4 Choice of Programming Language 59

## 4.5 Software and Hardware Requirements 59

## 4.6 Data Description 60

4.6.1 Models Vs Matrices: 60

## 4.7 Exploratory Data Analysis 64

# **CHAPTER FIVE**

# **CONCLUSION AND RECOMMENDATION**

## 5.1 Conclusion 67

## 5.2 Recommendations 68

Reference 69

**LIST OF FIGURES**

Figure 2.1 Statistics of Global rise of virtual Harassment 2004 -2020 (Camparitech, 2020) 18

Figure 2.2 Total number of active social media users in Nigeria from 2017 to 2022*(in millions)* (statista, 2022) 19

Figure2.3 Most popular social networks worldwide as of January 2022, ranked by number of monthly active users*(in millions)* (statista, 2022)

### Figure 3.1 An illustration of the ensemble learning method. *(Image by Author: Stacked ML Model) 32*

### Figure3.2: Flow Chart For The Text Classification 41

Figure3.3: Process Of Analysis 42

### Figure 4.1: Graphical representations of Evaluation Metrics against Machine Learning

### Algorithms and ensemble learning for Dataset 1 45

### Figure 4.2: Graphical representation of Evaluation Metrics against Ensemble Learning Models for Dataset 1 46

### Figure 4.3: Graphical representations of Evaluation Metrics against Machine Learning

### Algorithms and ensemble learning for Dataset 2 49

Figure 4.4: Graphical representation of Evaluation Metrics against Ensemble Learning Models for Dataset 2 51

Figure 4.5: Graphical representations of Evaluation Metrics against Machine Learning Algorithms and ensemble learning for Dataset 3 53

Figure 4.6: Graphical representation of Evaluation Metrics against Ensemble Learning Models for Dataset 3 54

Figure 4.7: Graphical representations of Evaluation Metrics against Machine Learning Algorithms and ensemble learning for Dataset 4 57

Figure 4.8: Graphical representation of Evaluation Metrics against Ensemble Learning Models for Dataset 4 58

Figure 4.9 Jupiter Notebook Interface 60

### Figure 4.10 The Comparative Analysis plot based on model Accuracy 62

### Figure 4.11 The Comparative Analysis plot based on the Dataset 62

### Figure 4.12 The Comparative Analysis graph based on models for each dataset. 63

### Figure 4.13: The evaluation matrices by the four datasets. 63

### Figure4.14 The Plot Average performance of evaluation Metrics by model 64

Figure4.15 The graph of word count for posts labeled Virtual Harassment and their

frequency 64

### Figure4.16 The graph of most common words labeled NOT-Virtual Harassment with frequency 65

### Figure4.17 Word cloud for random unlabeled text 65

Figure4.18 Word cloud for posts labeled to be Virtual Harassment text 66

### Figure4.19 Word cloud for posts labeled non-Virtual Harassment text 66

**LIST OF TABLES**

table 1: Summary of related articles 14

table 3.1: Dataset source 21

## Table 3.2: For comparative analysis showing the justification of the choice for Ensemble classifiers. 37

### Table 3.3: The Confusion Matrix 38

### Table 4.1 Result of Dataset 1 (kaggle, 2021) using the Machine Learning Algorithms and Ensemble learning methods 44

### Table 4.2 Result of Dataset 2 (YouTube, 2022) using the Machine Learning Algorithms and Ensemble learning methods 48

### Table 4.3 Result of Dataset 3 (Bayzick, 2022) using the Machine Learning Algorithms and Ensemble learning methods 52

### Table 4.4 Result of Dataset 4 (Aggression, 2022) using the Machine Learning Algorithms and Ensemble learning methods 56

### Table 4.5: Dataset Result Summary using 11 Model predictions across 12matrices 61

### Table 4.6: Confusion matrix for ten Classification 61

**Abstract**

Internet social media platforms have grown to be quite popular, enabling a wide range of online users to stay in touch with their friends and relatives wherever they are at any time. This has led to a significant increase in virtual crime from earlier periods to the present. Users are harassed online, either by publishing unpleasant or insulting comments about a specific user or by stealing personal information about that user and using it against them. Virtual harassment is referred to as an example of virtual crime that has posed a great threat to online social media users both mentally and psychologically. This project aims at comparing the traditional classification models and Ensemble learning in classifying virtual harassment in online social media networks by training both models using four different datasets. The classification models considered are: Naïve Bayes NB, Decision Tree DT, K Nearest Neighbour KNN, Logistics Regression LR, Neural Network NN, Quadratic Discriminant Analysis QDA and Support Vector Machine SVM, and four Ensemble learning model – Random Forest, Gradient Boosting, Ada Boost, and Max Voting. Finally, a comparison of the results was carried out using twelve evaluation metrics namely: Accuracy, Precision, Recall, F1-measure, Specificity, Matthew’s Correlation Coefficient MCC, Cohen’s Kappa Coefficient KAPPA, Area Under Curve AUC, False Discovery Rate FDR, False Negative Rate FNR, False Positive Rate FPR and Negative Predictive Value NPV were used to show the validity of our algorithms. At the end of the experiments, For Dataset 1, Logistics Regression had the highest accuracy of **0.6923** for machine leaning algorithms; Max Voting Ensemble had the highest accuracy of **0.7047**. For dataset 2, K-Nearest Neighbor, Support Vector Machine, and Logistics Regression all had the same highest accuracy of **0.8769** in the machine learning algorithm category while Random Forest (Manuel, Francisco, Victor, & Fidel, 2021) had the highest accuracy of **0.8779**. For dataset 3, the Support Vector Machine had the highest accuracy of **0.9243** for the machine learning algorithms while the Random forest ensemble had the highest accuracy of **0.9258**. For dataset 4, the Support Vector Machine and Logistics Regression both had **0.8383** while the Max voting ensemble obtained an accuracy of **0.8280**. Bar chart was used to graphically represent our results showing the minimum, maximum, and quartile ranges.

# **CHAPTER ONE**

# **INTRODUCTION**

## **1.0 Background to the study:**

Nowadays, billions of users are using various social media platforms such as Twitter, Facebook, Instagram, Pinterest, LinkedIn, YouTube, etc. as a rich source of data for mining information. Twitter and Facebook have become excellent initial points for social media analysis. Instagram is used majorly for the marketing of goods and services via short videos and write-ups, LinkedIn, on the other hand, is more of a professional handle where people share their work experience and other details for prospective employers to view and make recommendations while YouTube contains audio recordings and videos that sometimes serve as a learning or teaching aid, motivational talks, etc. Social network sites and micro-blogging sites are considered very good sources of information because people share their thoughts and discuss a certain topic freely (statista, 2022).

Analysis of information from various social media platforms has led to a variety of research. Examples include prediction of sales and the stock market, notification of events such as earthquakes, analysis of both natural and man-made disasters and public health information, estimation of public sentiment during elections and recession, fake news detection, and cybercrime identification such as Virtual Harassment. During an annual general sales meeting, for example, for a given medical laboratory with different branches across the country, such user-generated data can be very useful in projecting the general activities being carried out by each branch such as the number of visits by a particular patient, number of referrals from a particular region, the most frequent referrals, the aspect of medical evaluation with the most visit such as ECG and scan vs cardiovascular patients and products (drugs) sold by pharmacy and services rendered across each unit/region. Recently conducted surveys have revealed that such online reviews from the public and the company have played a very important role in the strategies of company establishment and expansion. Similarly, the information posted on these social media platforms either as an individual or a group as in the case of Aggression data is a huge amount of resource for obtaining the thoughts and opinions of the general public. The retrieval and analysis of such information in cases where a person or a group is been intimidated, insulted, or victimized is often referred to as **virtual harassment**. The term “Virtual Harassment or online bullying” cuts across several works or phases of life and it occurs at every face of human life.

virtual Harassment is seen in Education, social lifestyle, entertainment, politics, political class, workplaces or offices, technology, and computing advancements, etc. social media platforms have become a note of concern as it is used to oppress people at different stages, of different ages and works of life i.e. health, education, psychology, social life, etc. It cuts across all races, countries, ages, gender, families, and different groups. online crimes include Racism, Religious segregation, gender inequality, etc. Lots of research have been carried out to automatically detect Virtual Harassment and to classify the data extracted from virtual harassment by exploring social media platforms such as YouTube, Twitter, Facebook, Instagram, LinkedIn, Pinterest, and so on. One of the very common analyses which can be performed on a large number of data is the Classification of Virtual Harassment (Camparitech, 2020).

Virtual Harassment is becoming a popular way to classify the semantic orientation from the text, mainly because of the users of the social media handles who are free to express their feelings, thoughts, and impressions concerning a specific topic or an individual. Recent search technologies can effectively help users to obtain result data, which is related to their searched keywords. But, the semantic orientation of the content, which is more important, is the context in the reviews or opinions and it is not populated by current search engines.

For this given review, the classifier tries to classify the review into a positive category (virtual harassment) or negative category(non-virtual harassment), or neutral category(undefined).

One useful application of the classification of text (data extracted from social media networks) is that the transformed and highly intelligent data mining approaches now allow organizations to collect, categorize, and analyze users’ reviews and comments from micro-blogging sites regarding their services and products. This type of analysis makes those organizations capable to assess, what the consumers want, what they disapprove of, and what measures can be taken to sustain and improve the performance of products and services (Cambridge University Press, 2019).

However, reviews in natural language are usually expressed in subtle or complex ways which is difficult to analyze. So, the challenge of classification may not be overcome by simple text-categorization approaches such as n-gram or keyword identification methods. Hence, a better tool for automated classification of opinions, thoughts, messages, or sentiments to ascertain Virtual Harassment is to use machine learning (supervised learning) techniques that classify texts (data) into positive or negative aspects or polarity of public opinions.

Since the inception of the social media era, researchers have been motivated to investigate the capabilities, effects, and usefulness of these social media networks. These effects include fake news dissemination and detection, opinions, and sentiments, the most challenging is virtual harassment which has led to the application of artificial intelligence in network node vulnerability. Machine learning which is a sunset of Artificial intelligence is the study of data-driven methods capable of imitating, understanding, detecting, and assisting human and genetic information processing tasks. Many related issues arise such as how to collate data, sort data, compress data, interpret data, and process data. Often these methods are not necessarily directed to imitating human processing directly but rather to enhance data processing such as in predicting the outcome of an event rapidly. On a wider scale, machine learning and related fields aim to study something useful about the situation within which the agent functions. This agent can be software said to be an “Intelligent System”. It perceives its environment and takes necessary actions and still performs the task for which it was designed irrespective of the circumstance it faces. All of these functions are done autonomously i.e. taking decisions, thinking, etc. Machine learning makes more emphasis on using data to drive and adapt the model. Machine learning has its origin in the artificial intelligence field and its method is a member of statistical techniques. These modeling techniques are flexible enough to handle multifaceted problems with multiple instances, which are typically traditional approaches, making them ideal for modeling biological systems. Machine learning is turning data into information and acquiring intelligence from a particular set of data (Jason, 2020).

Retrieval of data set from a repository like Bayzick, Kaggle, Instagram, facebook, and YouTube respectively be done manually based on the huge volume of data available, there is a need to apply machine learning algorithms that help in classifying these datasets appropriately, the algorithms include Naive Bayes’, Random Forest, KNN (K-nearest Neighbor), Support Vector Machine and lots more. These algorithms have advanced the analysis of Virtual Harassment as part of the existence of a vulnerability in the structure, design, and architecture of computers and other smart electronic and computational devices, especially in the operating system, applications, and network structure. The identification and evaluation of the network node vulnerability in this case Virtual Harassment is a key issue in information security research and this has attracted more attention in recent times. Due to the availability of various algorithms for classifying data, there’s a need for measuring and comparing the performance of each algorithm with another to choose the best algorithm for a specific task which brings an accurate result from data classification for better management decisions, in this case, using these algorithms in Virtual Harassment detection. The vulnerability of network nodes is not just allied with the ability to resist intimidation but also influences the stable development of the network nodes in the long run.

## **1.1 Statement of the Problem**

The rate of Virtual Harassment on social media is characterized by intense vulnerability to the attack on network core nodes, therefore, Virtual harassment is a new form of Harassment that follows both students and working adults from the perimeters of their various schools and offices directly into their homes. It is a major problem right now among both old and young people (teenagers inclusive) who utilize social media platforms for communication and information sharing. The vast majority of victims of virtual harassment are being bullied from the moment they wake up to check the social media accounts and handles on their phone or email until they shut down their computers, laptops, phones, or any other computational device. This has had a significant negative impact on youth (young and old) and teenagers' emotions, psychological well-being, intimidation, low self-esteem, and interpersonal relationships because the so-called harasser can be anywhere, victimizing another person from the comfort of his own home, there is, therefore, need to take necessary action to help reduce and detect virtual harassment posts on social media platforms.

## **1.2 Aim and Objectives**

**1.2.1 Aim**

This project aims to evaluate ensemble learning models and traditional machine learning techniques in detecting virtual harassment using four (4) different dataset models

**1.2.2 Objectives**

* Experiment on the dataset by testing and training it using a machine learning approach with different seven types of classifiers and four ensemble learning
* Comparison and evaluation of machine learning Algorithms and ensemble learning models when it comes to detecting virtual harassment
* Making Pictorial, mathematical and statistical representations of the result and concluding how to detect virtual harassment in online social platforms

## **13 Contribution to Knowledge**

Four different social media datasets were used in the course of this project alongside 7 machine learning algorithms and 4 Ensemble learning models were incorporated into this project, statistical, mathematical, and pictorial representations were used to show the relationship that exists between the dataset.

## **1.4 Limitations of the Study**

This project focuses on detecting virtual harassment on four social media networks using a series of classifiers to validate our proposition, that the study of virtual harassment in social media Networks using machine learning and ensemble leaning is still in its infancy, there are still several other social media platforms that are prone to virtual harassment attack.

Indeed, the importance of this study cannot be over-emphasized. This study is important and beneficial to individuals, educational institutions, and corporate entities that are keen on making profitable decisions in their daily business transaction procedures or products and services or feedback from customers based on available data.

# **CHAPTER TWO**

# **LITERATURE REVIEW**

## **2.1 Virtual Harassment**

According to (uslegal), Online Harassment is referred to as "Virtual Harassment." The use of email, instant messaging, and offensive websites to bully or otherwise harass a person or group is known as Virtual harassment or Harassment. Flames, remarks made in chat rooms, the sending of rude or nasty emails, or even disturbing others by commenting on blogs or social networking sites are all examples of Virtual harassment. As the perpetrator of Virtual Harassment remains anonymous while threatening others online, it can be challenging to identify them.

Lexico Oxford Dictionary “Virtual harassment can be defined as the use of electronic communication to bully a person, typically by sending messages of an intimidating or threatening nature” (Lexico, 2019).

According to Dictiory.com “Virtual harassment can be defined as the act of harassing someone online by sending or posting mean messages, usually anonymously” (Dictionary, 2019).

According to Cambridge Dictionary, “Virtual harassment can be defined as the activity of using the internet to harm or frighten another person, especially by sending them unpleasant messages” (Cambridge University Press, 2019).

According to Meriam-Webster Dictionary, “Virtual harassment is the electronic posting of mean-spirited messages about a person (such as a student) often done anonymously” (Merriam-Webster, 2019).

## **2.2 History of Virtual Harassment**

### **2.2.1 Harassment - definition**

According to Richard (2012), the phrase "harassment" dates back to the 1530s. For harassment to take place, two people must be involved: "a tyrant" and the "tortured target." Abuse can be verbal, physical, audible, or through other ways, to assume dominance and power. These agreements may be direct (i.e., a declaration of defeat, verbal assault, text messages, online sexual harassment, sending nude pictures, etc.) or indirect (i.e. online comments, tales, chat, stalking, online assault response, etc.)­

### **2.2.2 Origins of Harassment**

All living things share a primitive need to continue to exist (both plants and animals). We all need to compete for the finite amount of attributes that the earth has to provide if we want to exist. Beginning on site, there was a constant drive to surpass others and overcome any obstacles in their path. This survival mindset has persisted from generation to generation as a result of our upbringing in a hostile environment.

This competitive and survival mindset has spread into the fields of academia, society, entertainment, and business. Even though the dominant spirit varies between cultures, ethnicities, tribes, religious and social groups, it is important to take into account the traditional order and the kind of laws that the ruling authority imposes. Success and hard labor go hand in hand, and the idea of hard effort can shape a country where harassment is unintentionally introduced as a means of survival at a very young age.

Some of the most common forms of virtual harassment are as follows (Sourabh & Vaivhav, Cyberbullying Detection and Prevention: Data Mining psychological perspective, 2014):

* **Sexual exploitation:** Sending of indecent nude pictures and dirty chats.
* **Flickering**: Heated online influences and competitions using bad-mannered and offensive linguistic language
* **Infuriation**: Constantly conveyance unpleasant, aggressive, or intimidating posts.
* **Slander**: Revealing the top-secret of an individual or blathers intending to hurt the name of a person.
* **Impersonation**: Breaking into the victim’s account and sending mails.
* **Deception**: Deceiving the victim with enlightening thoughtful information and conveying it to others.
* **Collaborative Gaming**: Most gaming consoles allow people to link up and play online enabling the chance to abuse through chats and comments

## **2.3 Traditional/Old-fashioned Harassment verses virtual harassment**

### **2.3.1 Traditional/Old Fashioned Harassment**

The traditional kind of harassment, which necessitates face-to-face interaction between the tormentor and the sufferer, can be separated from virtual harassment. From 2004 to 2020, the increase in virtual harassment is depicted graphically in Fig 2.1. Name-calling and embarrassing someone in public is examples of traditional forms of harassment. (Sophia, 2016). Females are more frequently the victims of traditional harassment compared to their male counterparts. It has to do with the first victim being verbally harassed. Another significant type of conventional harassment is malicious mocking, which may be overt or covert. Open harassment leaves visible marks that are visible to the victim and those around them, while covert harassment employs characters that are obscured or difficult to understand.



**Figure 2.1 Statistics of Global rise of virtual Harassment 2004 -2020** (Camparitech, 2020)

**2.3.2 The features of virtual harassment include**:

1. Virtual harassment takes place every minute of the day and every hour of the week due to the increased use of the internet and social network nodes and apps.
2. Awkward and humiliating messages about an individual can swiftly spread to an extremely large audience.
3. Virtual bullies do not have to appear before their victims, due to their secrecy making them feel “not guilty”
4. it is difficult to track down who has conceded out occurrences of virtual harassment,
5. conduct can be tedious or hostile
6. Sufferers of virtual harassment find it difficult to deal with bullies in actual life.
7. Victims of virtual harassment have pain cooperating socially and therefore have little or no friends. (Moreno & Kota, 2014)

The development of the human community is inversely correlated with technological advancement. The way people communicate with one another, however, has altered as a result of the internet. Because of the advancements made possible by the human race's development in various scientific and technological domains, atrocities of all kinds have become more prevalent and frequent. This is evident when looking at how traditional harassment, often known as "virtual harassment," has grown into a significant problem.

Although harassment and virtual harassment share similar definitions and skills, they also differ slightly in that virtual harassment allows the harasser to conceal their identity behind a system, unlike traditional harassment. Being able to remain invisible makes it easier to take advantage of the victim because they do not need to physically interact with the victim or wait for a response. The ability to complete any task from any location at any time without taking distance into account has given today's kids the capacity to commit wicked deeds.

## **2.4 A Technological Evolution**

With the advancement of science and technology, Harassment has multiplied, and the introduction of the internet and various social media platforms has provided a virtual platform for youth and teenagers to assault themselves (Nnamdimi & Sheeba, 2015). We have several online programs for communication purposes that enables youths and teenagers to spend several hours communicating with themselves privately (e.g. Messenger on facebook and Instagram, Whatsapp) or in chat room that are public (Twitter) to talk about the latest “gist”.

Technology is dynamic and we have experienced different innovations in the way we communicate. In the late 60s and early 70s cell phones were introduced and this has changed the manner people communicate with themselves (Shiels, 2003) Nevertheless, these telephonic devices weren’t used by youths and teenagers, not until the 1990s, this was when they when viral in the hand of the youth and teenagers. Nowadays, over 75% of teenagers now own a phone and they send over 8,000 text messages per month, although many parents presume they are buying the phones for their children to watch after them or probably for protective reasons which we know is not always true. Youth and teenagers nowadays acknowledge the fact that their phones are their best weapon for “virtual harassment” (Lenhart, 2019).

In addition, the advancement in Internet has allowed brought about more websites which brought about the different types of social media platforms we have. MySpace is a mostly deliberated social media innovator, the site enables users to create profiles and communicate directly in a virtual space with their friends and enemies in a similar way. Almost all our social media platform where a profile is being created requires you to provide some personal information as a requirement to own a profile such information includes name, date of birth, location, etc. Revealing such information at times can be dangerous because the profile is made public and anyone can access that information, which could be kept private if it were to be a face-to-face interaction. This weakness of the social media platform encourages many teenagers or youth to partake in virtual harassment actions (Nicole, 2009). Another weakness is that users can submit incorrect information to create a code-named profile which enables teenagers and youths to hide their real identities thereby giving them the privilege to write/post anything against another person without the need to be anxious about what follows or the side effect on the other person. Social media sites like Facebook, Twitter, and Google+ are liable to virtual harassment abuses.

## **2.5 Related Works**

Nureni et. al., in 2021 used Twitter datasets, as well as made effort to analyze well-known classification techniques and to suggest an ensemble model for detecting instances of cyberbullying. Naive Bayes, KT Nearest Neighbors, Logistic Regression, Decision Tree, Random Forest, Linear Support Vector Classifier, Adaptive Boosting, Stochastic Gradient Descent, and Bagging classifiers are some of the techniques used for evaluation. In experiments, the classifiers were compared against four metrics: accuracy, precision, recall, and F1 score. The outcomes show how each algorithm's performance compared to its relevant measures. Compared to the linear support vector classifier (SVC), the ensemble model produced better results of all. The medians for the Random Forest classifier across the datasets are 0.77, 0.73, and 0.94, making it the top-performing classifier. With medians of 0.77, 0.66, and 0.94 compared to the linear support vector classifier's 0.59, 0.42, and 0.86, the ensemble model has demonstrated an improvement in the performance of its constituent classifiers (Azeez et. al., 2021).

Manuel et. al., 2021 followed two supervised learning methods namely: Threshold and dual. Results show how to improve baseline detection models by up to 42%. Experiment with the dataset from some other social media platforms, Use a random forest for negative models, and an Extra tree for the positive models.

Celestine et. al., 2020 conducted an empirical analysis to determine the effectiveness and performance of deep learning algorithms in detecting insults in social media commentary. Results show that (Bidirectional Long Short-Term memory) BLSTM model achieved high accuracy and F1-measure scores in comparison to (Recurrent Neural Network)RNN, (Long Short-Term Memory)LSTM and (Gated Recurrent Units)GRU. Deep learning models can be most effective against cyberbullying when directly compared with others and paves the way for future hybrid technologies that may be employed to combat this serious online issue.

Quite a lot of methods have been proposed in the past year to quantity and sense unpleasant or insulting content and behaviors on Instagram (Hosseinmardi et. al., 2015), YouTube (Chen et. al., 2012), 4chan (Hine et al., 2017), Yahoo Finance (Djuric et. al., 2017), and Yahoo Answers (Kayes et. al., 2015).

Chen, et al., 2012 combined physical and written characteristics (such as the ratio of imperative sentences, verbs, adverbs, and adjectives as offensive words)to foresee a user’s ability in creating aggressive content in the comments made on YouTube, while Djuric, et al., 2017 depended on word implanting to distinguish foul comments on Yahoo Finance. Nobata, et al., 2016 Implemented hate speech recognition on Yahoo Finance and News data, using a classification based on supervised learning. Kayes, et al., 2015 discovered that users have a habit of flagging offensive content forwarded on Yahoo Answers in a tremendously correct way (as long-established by mortal annotators). Likewise, some users meaningfully diverge from public standards, posting a big amount of content that is flagged as offensive. After careful extraction of the features, they also showed that it is likely to use machine learning methods to predict which users will be suspended.

Dinakar, et al., 2011 sensed virtual harassment by disintegrating it into the detection of sensitive topics. Comments from YouTube were collected from contentious videos using manual comments to distinguish them and perform a “bag-of-words” driven text classification. Van Hee, et al., 2015learnt language features in virtual harassment-related content pulled out from Ask.fm, intending to detect fine-grained types of virtual harassment, such as intimidations and abuses. Besides the target and harasser, they also identified bystander-protectors and eye witness-assistants, who support, individually, the victim or the harasser.

Hosseinmardi, et al., 2015 studied pictures posted on Instagram and their related comments to detect and distinguish between virtual- aggression and virtual harassment. In conclusion, authors (Saravanaraj et. al., 2016) offered an attitude for detecting Harassment arguments found in twitter tweets, plus demographics about bullies(such as their age and gender). The aforementioned work frequently used structures(such as punctuation, URLs, part-of-speech, n-grams, Bag of Words (BoW), as well as lexical features depending on dictionaries of offensive words, and user-based structures such as user’s association period activity, amount of friends/followers, etc. Diverse supervised methods have been implored for detection: (Nobata et. al., 2016) used a regression model, whereas (Dadvar et. al., 2014), (Dinakar et. al., 2011), (Van Hee, et al., 2015) relied on other approaches like Naive Bayes, Support Vector Machines (SVM), and Decision Trees (J48). Dissimilarly, Hosseinmardi et. al., 2015, used a graph-based approach based on likes and comments to figure bipartite graphs and classify negative conduct.

A comparable graph-based approach is also used by Hosseinmardi et. al., 2015. Romanticism study of text can also add useful structures in detecting aggressive or insulting content. For instance, (Nahar et. al., 2012) used sentiment records of data collected from “Kongregate” (an online gaming site), “Slashdot”, and MySpace. They made use of a probabilistic sentiment investigation method to differentiate between bullies and non-bullies and rank the most important users based on a killer victim graph built from exchanged messages. Xu, Zhu, & Bellmore, 2012 depended on sentiment to categorize sufferers on Twitter who pose a high risk to themselves or others. Besides using positive and negative sentiments, they deliberated precise emotions such as anger, embarrassment, and sadness. Finally, Patch, 2015 studied the presence of emotions such as anger, sadness, and fear in Harassment occurrences on Twitter (Pieschl et. al., 2013)

Nandhini & Sheeba, 2015proposed a model that uses the Naïve Bayes machine learning approach and by their effort, they attained 91% correctness their dataset was gotten from MySpace.com, and then they projected an additional model (Nandhini & Sheeba , 2015) Naïve Bayes classifier and genetic operations (FuzGen) and they attained 87% accurateness. Another approach by Walisa, Lodchakorn, Pimpaka, Piyaporn, & Pirom, 2017 improved the Naïve Bayes classifier for removing the words and investigating loaded pattern gathering, and by this approach, they achieved 95.79% accurateness on datasets from “Slashdot, Kongregate, and MySpace”. Nevertheless, they had difficulties with the clustering process because it doesn’t work in parallel. Likewise, in the methodology proposed by (Shane et. al., 2018) the War of Tanks game cha was used t to collect their dataset and classified them manually and comparisons were made with the simple Naïve classification that makes use of emotional exploration as a characteristic, they had poor computation results when the dataset classified manually were compared with theirs.

Furthermore, Sani & Livia, 2017 proposed a method having gotten their dataset from kaggle, two different classifiers namely: Naïve Bayes and SVM were used. The Naïve Bayes classifier produced average correctness of 92.81% while SVM with poly kernel yielded an accuracy of 97.11%, but they did not reference the size of the dataset used for testing and testing, there is the possibility that their result might not be trustworthy. Another Approach by (Karthik, Birago, Catherine, Henry, & Roslind, 2012) that intended to distinguish obvious Harassment language relating to (1) Sexuality,(2) Race & Culture, and (3) intelligence, the data set used was gotten from YouTube comment, two different classifiers namely SVM and Naïve Bayes were used and generated results; SVM produces correctness of 66% while Naïve Bayes produces correctness of 63%

Moving on to (Michele et. al., 2016), projected a new method for detecting virtual harassment by implementing an unsupervised approach, they used the classifiers inconsistently over their dataset, using SVM on FormSpring and achieving67% on the ability to remember, applying GHSOM on YouTube and realizing60% exactness, 69% correctness and 94% remembrance, applying Naïve Bayes on Twitter and attaining 67% correctness.

Furthermore, Batoul, Maroun, & Ahmed, 2017 came up with a model to sense virtual harassment but with the use of Arabic language they used Naïve Bayes and got 90.85% accuracy, and with SVM they got 94.1% exactness but a rate of false positive was very high.

Another type of method using “Deep Learning and Neural Networks”, one out of the projected methods is in the paper by Xiang, et al., 2016. They used novel enunciation centered on a sophistication neural network, thereby lessening the difficulty related to noise and Harassment data scarcity to counter imbalance in the class. They got 1,313 messages from twitter, and 13,000 messages from formspring.me. They were unable to calculate the accuracy of the dataset gotten from twitter because they were imbalanced. They achieved 56%on exactness, 78% recall, and 96% accuracy, while achieving high accuracy their dataset was unbalanced, it, therefore, produced incorrect output and which was reflected in the score of exactness which was 56%. Chikashi, Joel, Achint, Yashar, & YI, 2016 displayed that using abusive language has increased recently, a framework called Vowpalwabb was used for their classification, and they also established a supervised classification methodology with NLP structures that outclassed the deep learning approach, The F-Score extended to 0.817 using dataset composed Yahoo News and Finance comment post.

Rui et. al., 2016 suggested a framework that was mainly for detecting virtual harassment, they made use of embedded words that are similar to insulting words weights were assigned to those words to obtain the features related to Harassment., and SVM was used as the main classifier and obtained correctness of 79.4%. (Sourabh & Vaibhav, Cyberbullying detection and prevention: Data mining and psychological perspective, 2014) Projected extra method, they collected their dataset from MySpace and marked them manually, then they used a Support Vector Machine classifier for their classification.

**table 1: Summary of related articles**

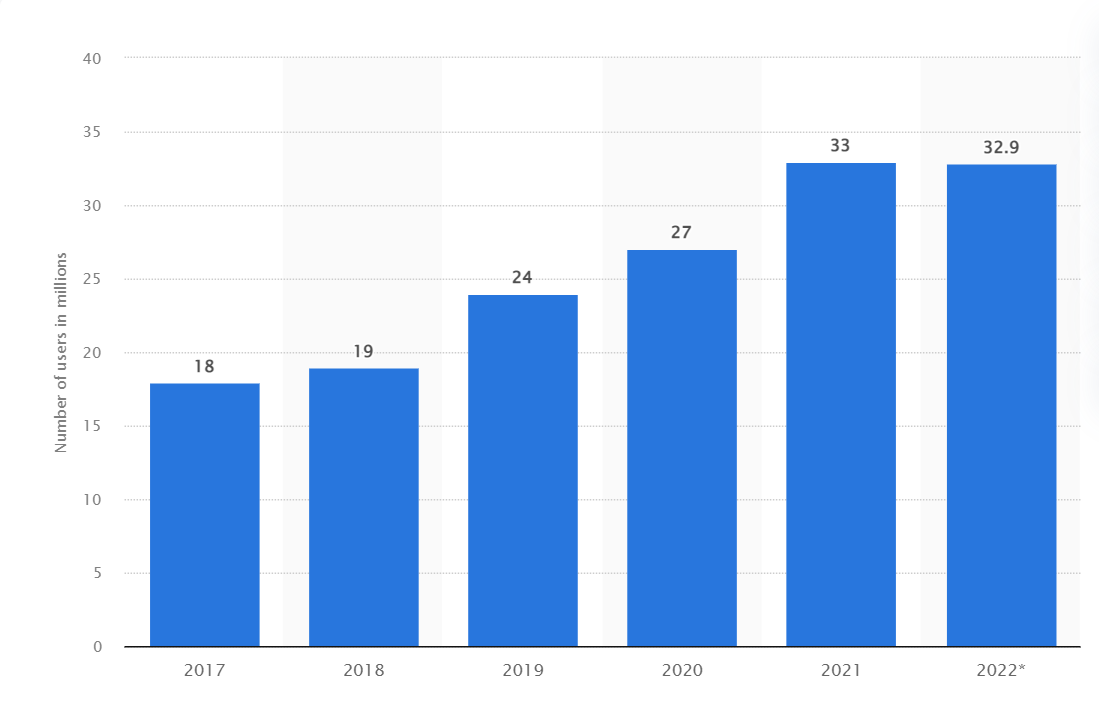
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Author** | **Year** | **Approach** | **Strength** | **Weakness** |
| (Nureni , Sunday, Chinazo, & Charles, 2021) | 2021 | Naive Bayes, KT Nearest Neighbors, Logistic Regression, Decision Tree, Random Forest, Linear Support Vector Classifier, Adaptive Boosting, Stochastic Gradient Descent, and Bagging classifiers are some of the techniques used for evaluation. | The medians for the Random Forest classifier across the datasets are 0.77, 0.73, and 0.94, making it the top-performing classifier. With medians of 0.77, 0.66, and 0.94 compared to the linear support vector classifier's 0.59, 0.42, and 0.86. | The ensemble model has demonstrated an improvement in the performance of its constituent classifiers |
| (Manuel, Francisco, Victor, & Fidel, 2021) | 2021 | Followed two supervised learning methods namely:   1. Threshold 2. dual | Results show how to improve baseline detection models by up to 42% | Experiment with a dataset from some other social media platforms.  Use random forest for negative models.  Extra tree for the positive models. |
| (Celestine , Gautam, Suleman, & Praveen , 2020) | 2020 | Empirical analysis to determine the effectiveness and performance of deep learning algorithms in detecting insults in social media commentary. | Results show that the BLSTM model achieved high accuracy and F1-measure scores in comparison to RNN, LSTM, and GRU. | Deep learning models can be most effective against cyberbullying when directly compared with others and paves the way for future hybrid technologies that may be employed to combat this serious online issue. |
| (Abaido) | 2019 | Enhance Timing Approach (ETA) and Ensemble learning. | SPSS was used for the reliability test and it showed satisfactory results for the research study (Alpha= .718) further results showed that Virtual Harassment exists on social media platforms at 91% positive. | Further quantitative research is required to assess the socio-psychological impacts of Virtual Harassment on victims in conservative societies |
| (Shane, William, Adrian, & Gordon) | 2018 | Datasets were gotten from the war of Tanks game and classifications were done manually. | It has a similarity with the Simple Naïve classification that uses emotional analysis. | The results produces  were very poor |
| (Walisa, Lodchakorn, Pimpaka, Piyaporn, & Pirom) | 2017 | Improved Naïve Bayes classifier was used to eliminate words and examine the loaded pattern | 95.79% correctness was achieved after the experiment | The cluster pattern doesn’t work in parallel |
| (Sani & Livia) | 2017 | Two classifiers were used Naïve Bayes and SVM and the data set was gotten from kaggle | 92,81% accuracy for Naïve Bayes and 97.11% for SVM | The dataset used for testing and training was not mentioned, hence their result isn’t credible |
| (Batoul, Maroun, & Ahmed) | 2017 | They made use of the Arabic language and the classifiers used were Naïve Bayes and SVM | 90.85% precision with Naïve Bayes and 94.1% precision on SVM | The result had a high rate of false Positive |
| (Michele, Emmanuel, & Alfredo ) | 2016 | An unsupervised learning approach was used | Accuracy of 67%, 60%,69%, 94% and 67% were achieved | The average levels of accuracy were low than when compared with supervised learning algorithms |
| (Xiang, et al.) | 2016 | Deep learning and Neural Networks approaches were used for the experiment | 56% exactness, 70% recall, and accuracy 96% | The data set was unbalanced while achieving high accuracy, so it gave incorrect output |
| (Rui, Anna, & Kezhi) | 2016 | They used words embedding makes a list of pre-defined words | 79.4% accuracy using Support Vector Machine | Only one classifier was used |
| (Chikashi, Joel , Achint, Yashar, & YI) | 2016 | Vowpalwabbit framework was used for classification and NLP features | It performs better when compared with the deep learning approach with about 81% accuracy | the other classifiers such as Naive Bayes gave better accuracy |
| (Nandhini & Sheeba ) | 2015 | Naïve Bayes machine learning Effort | 91% Accuracy was achieved | Efficiency is reduced when tried with another classifier |
| (Krishna & Narendra) | 2015 | Automated system for detecting virtual harassment text and images | The accuracy level of detection worked | The system needs to be thoroughly trained to successfully detect virtual harassment pictures in the future |
| (Karthik, Birago, Catherine, Henry, & Roslind) | 2012 | The data set gotten from the YouTube Comment section using two classifiers SVM and Naïve Bayes | SVM gave 66% Correctness and Naïve Bayes gave 63% correctness | The accuracy level given in the result was too low. |

[Vadivukarassi M. (2017)](#_ENREF_10) Connect to twitter and search for tweets that contain a particular keyword and evaluate the polarity of the tweets as positive or negative. They used Twitter API (Application Programming Interface) and the executed raw data are preprocessed using the natural language toolkit techniques and the chi-square is used for feature selection. They apply the Naïve Bayes classifier for training and testing the features which bring about sentimental polarity. The Naïve Bayes classifier is used to predict the probability for a given word to belong to a particular class either positive or negative based on the dictionary methods of the score.

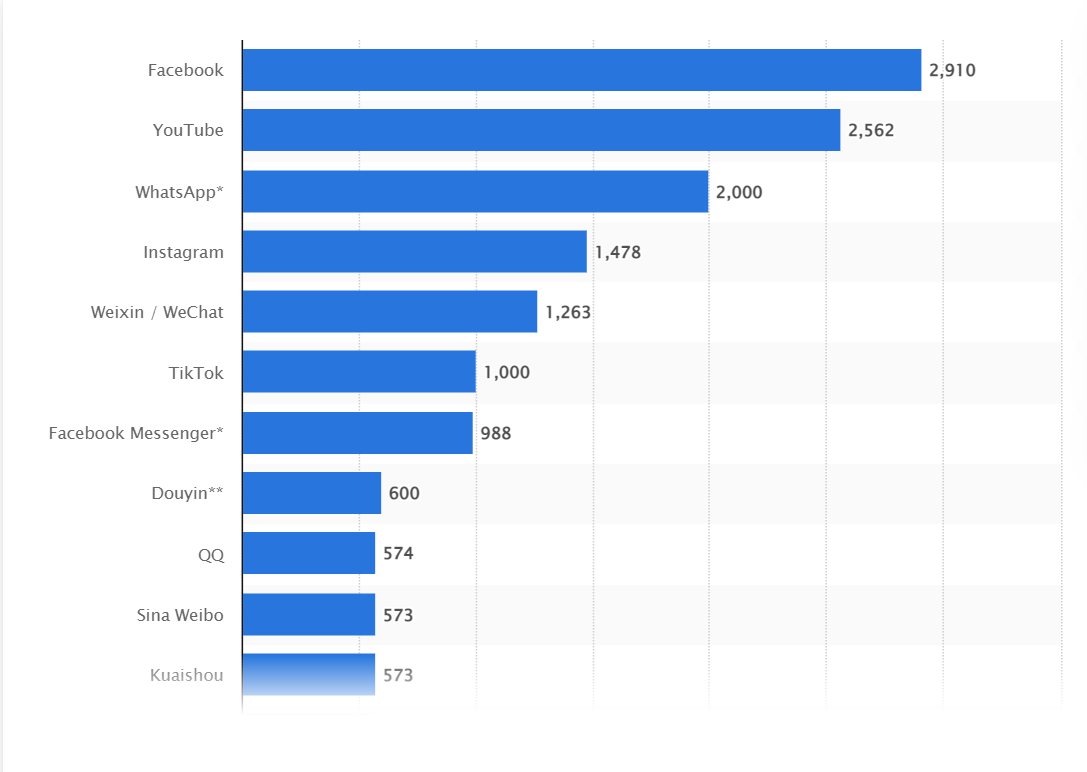
[Jasmine Norman in 2017](#_ENREF_8), performed virtual harassment to examine the percentage of people who have supported the demonetization and Indian budget 2017, they used the machine learning algorithm Naive Bayes classifier. They use the classifier algorithm (Naive Bayes) to classifier the tweets into positive and negative classes separately. With the help of the classifier, the feelings and estimation of the general population about the government's call to demonetization and its outcome on the proposed budget in 2017 were known.

[Mohammad S. H, 2012](#_ENREF_12) and [Feldman, 2013](#_ENREF_6) introduced sentiment classification problems in different levels which are document level, sentence level, word level, and aspect level. Some techniques that have been used to solve the problem of Virtual Harassment were introduced and discussed which have argued that Support Vector Machines (SVM) are more appropriate for sentiment classification than generative models because they can better differentiate mixed sentiments which involve both positive and negative words used in the same review. They suggested that when the set of training data is small a Naïve Bayes classifier might be more appropriate since SVMs must be exposed to a large set of data to build a high-quality classifier. [Munir Ahmad, 2018](#_ENREF_14) carried out a comparative analysis of Support Vector Machine (SVM) and other classifier algorithms.

[Abinash Tripathy, 2015](#_ENREF_2) attempted to classify Virtual Harassment for movie reviews using machine learning techniques. Two different algorithms Naive Bayes (NB) and Support Vector Machine (SVM) are implemented. It is observed that the SVM classifier outperforms every other classifier in predicting the sentiment of a review.



**Figure 2.2 Total number of active social media users in Nigeria from 2017 to 2022*(in millions)*** (statista, 2022)

****

**Figure 2.3 Most popular social networks worldwide as of January 2022, ranked by number of monthly active users*(in millions)***(statista, 2022)

# **CHAPTER THREE**

# **METHODOLOGY**

## **3.0 Data Collection**

### **3.0.1 Dataset 1**

The First dataset was extracted from Kaggle, to extensively make a study among LA and Boston youths who are between (18- 40 years). The Kaggle\_Parsed\_dataset consists of a total of 8,800 (eight thousand, eight hundred) tokens that are divided into chats where the youths raised concerns about the presidency and 9/11. The dataset was gotten from the Github website (GitHub, 2022). The extracted zip file dataset for aggression contained extensive information publicly available such as their user name and their public post - <https://github.com/jo5hxxvii/cyberbullying-text-classification>. The dataset was used by Sprugnoli, Menini, Tonelli, Oncini, & Piras, 2018 in “Creating an aggression Dataset to study pre-teen cyber harassment. The Kaggle dataset was used to train our machine learning and ensemble learning models respectively.

### **3.0.2 Dataset 2**

The Second dataset was gotten from the YouTube website; <https://github.com/jo5hxxvii/youtube_parsed_dataset-text-classification> which includes a large number of YouTube comment pages publicly made available on the site made based on the topic of random shootings by teenagers and young adults between the age of 14 and 56 years. The dataset contains text that may be considered vulgar, abusive, disrespectful, threatening, and suicidal. The dataset was used to check Toxic comment classification challenges. The dataset has been labeled by human raters for toxic behaviors. The toxicity is – toxic, severe toxic, obscene, threat, racist, racism, tribalism, and hate. The dataset was used to train the machine learning algorithms and ensemble learning models.

### **3.0.3 Dataset 3**

Dataset 3 was also gotten from the Bayzick website - <https://www.bayzick.com/bullying_dataset> and it can be accessed via “cyberbullying-text-classification-main/BayzickBullyingData/HumanConcensus” used for an online toxic comment classification challenge used to identification and classification of online toxic comments. Dataset 3 is another set of a large number of twitter comments which has been labeled by humans for toxic behaviors, sexism, and racism. The dataset contains – train.csv, test.csv, sample-submission.csv, and test labels.csv. the dataset was then used to train our machine learning and ensemble learning models in the course of the project.

### **3.0.4 Dataset 4**

The Second dataset was gotten from the Kaggle website; - <https://github.com/jo5hxxvii/aggression_parsed_dataset-text-classification> which includes a large number of 115,863(one hundred and fifteen thousand, eight hundred and sixty-three) comments. Rabbinic/pharisaic Judaism talks in form of comments publicly made available on the site made available for improvements to the current model to help online chats become more productive and respectful. The dataset contains text that may be considered as religious biases, tribalism, and racism. The dataset has been labeled by human raters for toxic behaviors. The dataset was used to train the machine learning algorithms and ensemble learning models.

### **table 3.1: Dataset source**

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | URL | Source | Remark |
| Dataset 1 | <https://github.com/jo5hxxvii/cyberbullying-text-classification> | Kaggle | GitHub, 2022 |
| Dataset 2 | <https://github.com/jo5hxxvii/youtube_parsed_dataset-text-classification> | You-tube | GitHub, 2022 |
| Dataset 3 | <https://www.bayzick.com/bullying_dataset> | Bayzick | Bayzick.com |
| Dataset 4 | <https://github.com/jo5hxxvii/aggression_parsed_dataset-text-classification> | Kaggle | GitHub, 2022 |

## **3.1 Data Preprocessing**

Data Pre-processing is an essential part of text classification and this is done because the datasets contain a large amount of vague information which needs to be eliminated. In pre-processing step, Data Cleansing, Integration and Transformation is firstly done, where all the special characters used like (! @ , . / \ ; : “ “\ [ ] { }… ) and the unnecessary blank spaces are removed. The CountVectorizer module in python is used to remove the English stop words from the dataset. Data transformation involves transforming or consolidating the data into formats that are suitable for mining. The following are examples of data transformation: **Smoothing**, for example, aims to eliminate noise from data and these methods include regression, grouping, and binning. **Aggregation** is the process of applying summary or aggregation operations to data. To calculate the amount of vitual harassment comments by a particular age bracket. This procedure is frequently used when building a data cube for the examination of data at various granularities. **Dimensional Reduction:** the process of identifying and removing unimportant, tangentially relevant, or duplicate features or dimensions. Data transformation procedures change the data into mining-ready formats. For instance, attribute data may be normalised to fall between narrow ranges, such 0 and 1.0. An inbuilt function called Vectorization was used and this process is shown in Figure 3.5, i.e. after cleaning the dataset, features can be extracted from it. The features are tokenized word of the dataset. These words need to be converted to numerical vectors so that each data can be represented in the form of numerical data. For example, a text that is positive for virtual hassraassment is assigned one “1” while a text that is negative or does not contain virtual harassment is assigned zero “0” as shown in Figure 3.6. **Data Reduction** is another process. It involves the creation of a Feature. Create new attributes that are considerably more effective than the existing ones in capturing the crucial facts in a data set. Threre are broad approaches to data reduction: Initial is the Feature Extraction, second is level specific, Data Mapping to New Space, Feature Development and Integrating features. All of these was achieved by using keywords in virtual harrsemment like black monkey, cunt, fag, bullock, etc. (Jasdeep Singh Malik, Prachi Goyal, Akhilesh K Sharma, 2010).

## **3.2 Featured Engineering**

By being familiar with the requirements of machine learning algorithms and generating the ideal input dataset, feature engineering aids in improving the effectiveness of machine learning models. A model for identifying virtual harassment was built using the features that were generated from the data provided by the comments and combined with the supervised machine learning framework. The text classification methods using the ML approach can be roughly divided into supervised and unsupervised learning methods ([Hao Wang, 2012](#_ENREF_7)). Supervised learning is an important technique for solving classification problems. Training the classifier makes it easier for future predictions for unknown data. The supervised methods make use of a large number of labeled training documents and this set of problems is known as the supervised method because we’re telling the algorithm what to predict. The training data includes the input and the desired results([Pollyanna Gonçalves, 2014](#_ENREF_15)). The Machine Learning Approach (ML) applies the famous ML algorithms which are Decision Tree, k-Nearest Neighbors, Logistics Regression, Naive Bayes, Neural Network, Quadratic Discriminant Analysis, and Support vector machines all of which use linguistic features.

The unsupervised methods are used when it is difficult to find these labeled training documents. The model is not provided with the target outputs during the training. These types of machine learning algorithms take the unlabeled input data and then with the help of different algorithms hidden pattern or structure is discovered. It does not consist of a category and they do not provide the correct targets at all and therefore rely on clustering. It can be used to cluster the input data in classes based on their statistical properties only. Examples of unsupervised algorithms include clustering (grouping similar items together) and the k-means clustering algorithm.

### **3.1.1 Activity Features**

The user's activity feature focuses on how frequently a user logs on, how frequently he or she publishes comments in the group, and how active the user is on the online social media platform. Over a year, researchers were able to conclude that more active individuals are more likely to encounter online bullies. (Balakrishnan, 2015)

### **3.1.2 User Feature**

A client's propensity for virtual harassment can be better understood by looking at the characteristics of an online social media site, including their age, gender, and personality. In general, virtual bullies are more aggressive and frequently have mental health issues, making them moodier, angrier, and tenser than average individuals. This might be seen in their comments on a particular post. (Mahmud et. al., 2013). However, they suggested that one's essay and blog writing online are somehow related to the personality of the user. Users' gender is one of the most important elements in determining whether virtual harassment is occurring in online communication networks; nonetheless, (Liu and Ruths, 2013) projected that men are more likely than women to engage in virtual bullying. Although the Aggression User Interface does not permit providing gender information, the classifiers were trained to recognize virtual harassment using the user's first name.

### **3.1.3 Content Features**

The use of rude or offensive language in an online communication platform can be utilized to spot virtual harassment because it signals hostility, rudeness, and aggressive behavior. A conversation may be categorized as harassment chat if it contains an offensive or profane word. The research was made by (Wang et. al., 2014) for detecting and identifying cursing and offensive behavior on the twitter Network. The dictionary (Profane) was been used to measure the number of profane/ Harassment related words in a particular post.

**3.3 Machine Learning Algorithms**

In this research work, seven (7) different Algorithms (classifiers) and four (4) different Ensemble Learning methods were used in this research while training the data set in the course of the project. Those machine learning Algorithms include:

1. **Decision Tree algorithm –**

**Decision trees** are easier to understand than their random forest counterpart, which synthesizes numerous decision trees into a single model and may be more effective for multi-class classification and other challenging artificial intelligence problems. They belong to the family of supervised learning **algorithms**. Problems are solved using **tree** representations, each internal node of the **tree** corresponds to an attribute, and each leaf node corresponds to a class label. There are several significant distinctions between decision trees and neural networks, even though they both offer distinct methods for classifying (or grouping) data into clusters that have similar traits (or features). To put it simply, decision trees perform best in straightforward situations with few variables, but neural networks excel in situations where the data includes complicated correlations between characteristics or values (i.e., is "dense") and decision trees don't. Decision trees are so frequently employed in modest data science projects as the first-line categorization technique. When dealing with a lot of high-dimensional data, however, they might not scale effectively, i.e. The decision tree construct classification models having the structure of a tree, the data set is then broken down into smaller and smaller subsets while a linked decision tree is developed incrementally. The classification or decision is represented by the leaf node while the node has two or more branches (Rajbanshi, 2021).

……………………………………………………………….3.1

…………….……..………..…………..3.2

Where, E – entrophy is a logarithmic algorithm for expolartion analysis and

G – gini for false reduction classification.

Eqa. 3.2, is the splitter (Pradhan, 2022).

1. **K-Nearest Neighbor (KNN) –**

One of the simplest categorization techniques is the K-nearest neighbors approach. It does a proximity comparison between every example of a given class and every other example of that class.

This is helpful in two ways:

1. It's simple to comprehend and use.

2. It is remarkably accurate.

We can just take the mean of each class, which is simple to calculate, making it simple to put into practice. Since KNN is not sensitive to the number of attributes that can be employed, it can also be used to solve classification issues involving several classes.

A parametric classification method is called K-Nearest Neighbor. An object is categorized based on the majority vote of its neighbors and is then put into the class with the highest percentage of support among its K closest neighbors. (K is a user-defined constant that is often a tiny, positive integer.) Data points are converted into feature vectors to execute KNN, which has its foundation in mathematical theories. (or their arithmetic equivalent). In other words, the algorithm determines the distance between the spots based on their mathematical values. Euclidean distance is a popular distance metric for continuous variables.

……….………3.3

= ………………………..………….………3.4

N denotes the number of samples in your training dataset, and k is equal to sqrt(N)

(Sreekesh, 2016).

1. **Logistics Regression** –

We can manage classes by using logistic regression, which is a non-linear extension of linear regression. This is accomplished by categorizing predictions according to a probability threshold. Think at the following illustration of the likelihood of passing an exam versus the number of hours studied. Assume that the variables Y and X indicate the number of hours spent studying and the likelihood of passing an exam, respectively. In that situation, we can use a regression predictor to fit a line through these points. Depending on how near the point's line is to a threshold, we can then determine if it is a pass or a fail. Even though this is a straightforward illustration, logistic regression is frequently utilized in practical applications, including multi-label classification difficulties and estimating creditworthiness across a range of categories. The naive Bayes classifier, which applies Bayes' theorem and produces a larger bias but lower variance, can be replaced with logistic regression (Jason, 2020).

* We know the equation of the straight line can be written as:

…………………………………………………3.5

* In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

……………………………………………3.6

* But we need a range between -[infinity] to +[infinity], then take the logarithm of the equation it will become:

……………………………………...….3.7

The above equation is the final equation for Logistic Regression.

Three different types of logistic regression can be distinguished based on the categories:

* Binomial: In a binomial logistic regression, the dependent variables can only be of one of two conceivable types, such as 0 or 1, Pass or Fail, etc.
* Multinomial: In multinomial logistic regression, the dependent variable may be one of three or more possible unordered kinds, such as "cat," "dogs," or "sheep."
* Ordinal: In ordinal logistic regression, the dependent variables can be categorized into one of three potentially ordered classes, such as "low," "Medium," or "High."

The algorithm measures the relationship between one or more independent variables which are the features obtained from the dataset and the dependent variable which are the labels that are to be predicted by calculating probabilities using the logistic function also known as the sigmoid function. The values obtained are then converted into binary values to deduce a prediction. The sigmoid function is an S-shaped curve that can acquire any real-valued number and map it into a value between the range of 0 and 1 but never exactly those limits.

*1 / (1 + …………*……………….………………………..3.8

Where e is equal to the base of the natural logarithm (Euler’s number) and value is equal to the actual numerical value to be transformed. (Brownlee, 2018)

1. **Naïve Bayes**:

Naive Bayes classifiers are a subset of linear classifiers where the assumption is that the value of a particular feature is independent of the value of any other feature. This means that we can use Bayes’ theorem to calculate the probability of a particular label given our data by just looking at each feature individually, without considering how features may interact with each other. Naive Bayes classifiers are often used in text classification because it’s easy to calculate probabilities from frequencies, and text typically has a large number of features (e.g., individual tokens in words). They are also popular in spam detection because they can deal with the high dimensionality of email data (e.g., all the different words used in an email) without overfitting the data.

…………….……………………………………3.9

This algorithm adopts the Naive Bayes algorithm, which is designed for training and classifying data that is dispersed according to the multi-variated Bernoulli distributions (i.e., even if there are several features, each one is presumptively a binary value (Boolean, Bernoulli) variable). The class expects samples to be represented as binary-valued feature vectors; otherwise, it will binarize the inputs by the parameters provided if a different form of data is provided. In addition, the Bernoulli distribution is an independent probability function in which a random variable can take one of two potential values—1 for success or 0 for failure.

Given a class variable or hypothesis (y) and a dependent feature or evidence (x1 – xn). Therefore,

*=* ……………………………... 3.10

where:  *are labels*

*are comments*

How probable was the hypothesis(labels) given the observed evidence(comments)?

is How probable is the evidence, given that the hypothesis is true?

is How probable was the hypothesis before observing the evidence?

is How probable is the new evidence under all possible hypotheses?

(Zhang, 2023)

1. **Neural Network (NN):**

One of the most popular machine learning algorithms has been artificial neural networks. They simulate biological brain networks in computers, as their name implies. More parameters are needed for more sophisticated models, which might make them slower than simpler techniques like logistic or linear regression algorithms in classifying new data points. However, because of their adaptability and scalability, they can easily handle enormous amounts of unlabeled data. A collection of parameters is trained on data using ANNs. The model's output, which could be either an input or an action, is then determined using these parameters. For instance, if we have a model for determining a person's likelihood of purchasing a car based on past purchases and demographics, To help it make future predictions, we can provide new data. Age, marital status, income, and other factors are just a few of the inputs we might utilize.

A neural network's nodes are linked together by weighted links according to how important they are to the result, which is commonly determined using the backpropagation error-minimization technique. The neurons that make up each layer of the network generally correlate to particular characteristics or qualities. The perceptron is the most basic type of artificial neural network. Deep neural networks are now employed to solve more challenging issues, such as image categorization, where there are numerous potential inputs (e.g., different road conditions for a self-driving car). The purpose of training is to fine-tune each node's parameters (weights) to best reflect the data it has seen. All nodes should be able to make judgments based on their parameters once they have all been taught (Rendyk, 2022).

The neural network equation looks like this:

……………………………….3.11

where,

* Z is the symbol for denotation of the above graphical representation of ANN.
* W is, are the weights or the beta coefficients
* X is, are the independent variables or the inputs, and
* Bias or intercept = W0

The fact that ANNs are so reliant on the data used to train them is one of their disadvantages. Overfitting may occur if the data are not indicative of the phenomenon we're attempting to forecast. By using regularization techniques, Akkio, 2022 reduces these problems. When compared to methods like linear regression, they also tend to be relatively slow, which could affect how well they work when training big models. With a model training procedure that is 100x faster than the competition, Akkio combats this. In other words, creating an artificial neural network only takes a few minutes. Another problem is that neural networks might be challenging to convey to others who aren't as knowledgeable about the subject. However, there is no other AI system has been able to achieve such rapid growth.

1. **Quadratic Discriminant Analysis:**

The LDA variation known as quadratic discriminant analysis (QDA) enables non-linear data separation. This is accomplished by using a quadratic curve as opposed to a linear boundary to match your data. Due to the quadratic operation required to determine the within-class variance for each class, QDA is more computationally demanding. However, if you have a large amount of training data and you think that the classes in your data are not linearly separable, QDA might be a better option than LDA (DataSklr, 2020).

………………………………………………………………………...3.12

1. **Support Vector Machine (SVM)**

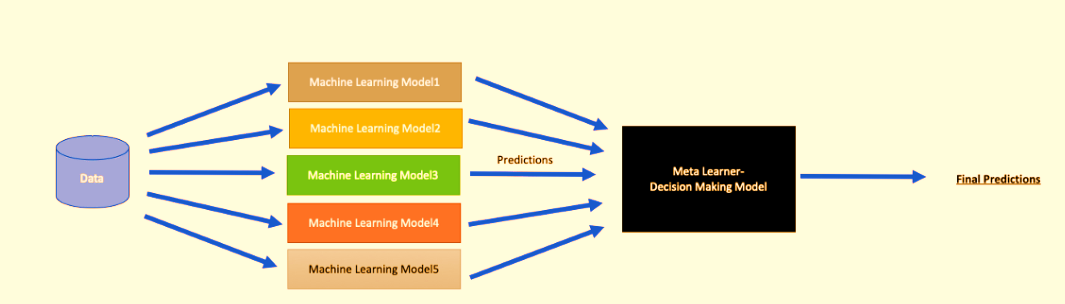
Support Vector Machines (SVM), which convert your data into a linear decision space, are reliable and efficient machine learning methods. After that, the algorithm chooses the best hyperplane in this linear decision space to divide your training data into distinct classes, such as valid and invalid emails. SVM excels at distinguishing related objects. As an illustration, all the samples in one group might be comparable, whilst certain examples in another group might be vastly dissimilar from one another. This makes it perfect for activities that can be classified effectively using a straight line, which is not the case for all classification problems, as you may imagine**.** It is a supervised learning model with related learning algorithms that provides data analysis for classification. The data are represented as points in space, mapped so that samples from the various categories are separated by as wide a gap as possible. The best hyperplane that separates the two classes in the optimal technique is found to classify the data into distinct groups. Support Finding the "hyperplane we x=0" that maximizes the margin between the two classes—which may be done by solving a quadratic objective function—allows Vector Machine to distinguish between positively and negatively labeled data (Zhang, 2023).

………………………………….…………………………….3.13

**3.4 Ensemble Learning**

The suggested ensemble model is a method that combines various machine learning classifiers and models to outperform the individual models. On the dataset, each component classifier is trained to make predictions. A final prediction is then created by combining these predictions. There are several ways to reach this conclusion, including stacking, voting, bagging, and boosting. Voting is employed in this study to determine the outcome. Here, the majority rule is implemented through the use of projected class labels for voting. The ensemble uses the constituent estimators' Multinomial NB, Linear SVC, and Logistic Regression (Azeez et al., 2021).

An individual learner is often trained by an existing learning algorithm, such as the C4.5 algorithm or the BP neural network algorithm, in the normal process of ensemble learning, which involves training a group of individual learners first and then combining them. If all of the individual learners in an ensemble belong to the same kind, it is said to be homogenous. For example, a "decision tree ensemble" would only contain decision trees, but a "neural network ensemble" would only contain neural networks. The individual learners are known as base learners for homogeneous ensembles, and the corresponding learning algorithms are known as base learning algorithms. In contrast, there is no single basis learner or base learning algorithm in a heterogeneous ensemble, which consists of several individual learners and learning algorithms. The results show that the term component learners or just individual learners are unusually used to describe individual learners in diverse ensembles. (Springer, 2021).



### **Figure 3.1 An illustration of the ensemble learning method. *(Image by Author: Stacked ML Model)***

An ensemble method or ensemble learning algorithm consists of aggregating multiple outputs made by a diverse set of predictors to obtain better results. Formally, based on a set of “weak” learners we are trying to use a “strong” learner for our model. Therefore, the purpose of using ensemble methods is to average out the outcome of individual predictions by diversifying the set of predictors, thus lowering the variance, to arrive at a powerful prediction model that reduces over-fitting our training set. The generalization capacity of an ensemble, which is made up of numerous learners, is frequently significantly stronger than that of an individual learner. This is particularly true for weak learners. Base learners are sometimes referred to as weak learners since theoretical research on ensemble learning frequently concentrates on weak learners. Although an ensemble of weak learners can theoretically produce a good performance, in practice individuals still favor strong learners for a variety of reasons, including a reduction in the number of individual learners and the ability to reuse previously learned information about the strong learners. It seems to sense that combining multiple traits will result in a product that is worse than the best while also being better than the worst. (Springer, 2021)

In the course of this project, 4 Ensemble learning Algorithms were used

1. **Random Forest Model**

(Dipayan Sarkar)In our case, a Random Forest (strong learner) is built as an ensemble of Decision Trees (weak learners) to perform different tasks such as regression and classification. Random Forests are the learning method for classification and regression. It constructs several decision trees at training time. To classify a new case, it sends the new case to each of the trees. Each tree performs classification and outputs a class. The output class is chosen based on majority voting which is the maximum number of similar classes generated by various trees that are considered the output of the Random Forest. Random Forest is an ensemble learning algorithm for classification. Random Forest generates a multitude of decision trees classified based on the aggregated decision of those trees. After training, Predictions for an unseen sample can be made by averaging the predictions from all the individual classification trees on

= ()………………………………………..……………………………3.14

The Random Forest algorithm does the selection of observations and features randomly to build several decision trees and then computes the average of the results. The random Forest Algorithm creates a random subset of the features and builds smaller trees using the subset created. Furthermore, Random Forest produces high accuracy through cross-validation, handling missing values, and maintaining the accuracy of a large proportion of data is another key. Random Forest classifiers don’t allow over-fitting trees into the model in case there are no more trees.

……………….…….………………..3.15

Random Forests are easy to learn and use for both professionals and laypeople with little research and programming required. It can easily be used by persons that don’t have a strong statistical background.

Decision Tree T3

Decision Tree T2

Decision Tree T1

Random Forest

A decision tree is made of a directed series of decisions, based on input variables value, and culminating in a classification of the target variable.

1. **Max Voting**

One of the easiest methods for integrating predictions from many machine learning algorithms is max-voting, which is typically used for classification tasks. Each base model makes a prediction and casts a vote for each sample in max-voting. The final prediction class only consists of the sample class that received the most votes.

………………………….……………………………..3.16

……………….………………..………….……………3.17

How is voting conducted? Voting is effective because the vote on an individual has less significance than the view of the majority. When the models may vote on discrete possibilities, maximum voting is utilized. The choice is the proposal that has received the most votes. It is employed to solve classification issues. The option with the most votes is chosen. Each machine learning model casts a vote.

3. **Ada Boost:**

**AdaBoost Classifier:** This is an ensemble classifier (consists of various classifier algorithms whose result is the combined output of the other classifier algorithms). The classifier combines weak classifier algorithms to produce strong classifier algorithms with the selection of training set at each iteration level and the designation of the correct amount of weight in the final voting. Any machine learning classifier algorithm can be the base classifier if it can accept the designation of weights so the training set. Two conditions are required for this classifier:

1. Training of the classifier iteratively on various weighed training sets
2. In each iteration, an accurate fit for the samples should be provided to mitigate errors in training.

AdaBoost Classifier works in the following steps:

1. Selection of a training subset randomly
2. Iteratively trains the AdaBoost model by choosing the training set based on the correct prediction of the previous training.
3. The largest weight is assigned to the incorrectly classified observation to ensure that these observations will have the highest probability of classification in the next iteration.
4. According to the accuracy of the classifier, the weight is assigned to the trained classifier in each iteration. The highest weight is assigned to the more accurate classifier.
5. The process undergoes iteration till the entire training data fits without any error or until the specified maximum number of estimators is exceeded.
6. For classification, perform a vote across all of the learning algorithms built. (Avinash, 2018)

Let's look at the mathematical formula and parameters.

………………………….....…………3.18

1. **h\_t(x)** is the output of weak classifier t for input x
2. alpha\_t is the weight assigned to the classifier.
3. alpha\_t is calculated as follows:
4. alpha\_t = 0.5 \* ln( (1 — E)/E): weight of the classifier is straight forward, it is based on the error rate E.

Initially, all the input training example has equal weightage.

Then, updating the training examples' weight, Following weak classifier training, we update each training example's weight using the formula below.

………………..………………………..3.19

Weight at the preceding level is D\_t.

By dividing each weight by the total weights, Z\_t, we normalize the weights. For instance, if the sum of all the estimated weights was 15.7, we would divide each weight by 15.7 to get 1.0 as the result.

For simplicity, y\_i is the y par of the training example's (x\_i, y\_i) y coordinate.

**4. Gradient Boosting:**

Classification using gradient boosting.

This approach allows for the optimization of any differentiable loss function and constructs an additive model in a forward stage-wise manner. Each stage involves fitting n classes\_ regression trees on the loss function's negative gradient, such as a binary or multiclass log loss. In the particular scenario of binary classification, just one regression tree is generated.

A considerably quicker version of this approach for the processing of intermediate datasets (n samples >= 10\_000) is sklearn.ensemble.HistGradientBoostingClassifier

………………………………………………………………………………3.20

where N is the total number of samples, N\_t is the number of samples at the current node, N\_t\_L is the number of samples in the left child, and N\_t\_R is the number of samples in the right child.

N, N\_t, N\_t\_R, and N\_t\_L all refer to the weighted sum if sample\_weight is passed.

## **Table 3.2: for comparative analysis showing the justification of the choice for Ensemble classifiers.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S/No** | **Random Forest** | **Max Voting** | **Ada Boost** | **Gradient Boosting** |
| 1. | Another ensemble machine learning algorithm that uses the bagging method is Random Forest. | For categorization issues, the max voting approach is typically utilized. | An additive model in which high-weight data points are used to highlight the flaws of earlier models. | A model that is additive and uses the gradient to identify the inadequacies of earlier models. |
| 2. | In contrast to the bagging meta estimator, random forest chooses a set of features at random, using those characteristics to determine the optimum split at each decision tree node. | With this method, predictions are made for each data point using a variety of models. | As decision stumps, the trees are typically cultivated. | The trees are typically grown to a larger depth, typically between 8 and 32 terminal nodes. |
| 3. | We produce random subsets from the original dataset (bootstrapping).  Only a random collection of features is taken into account at each node of the decision tree to determine the best split. | Each model's predictions are regarded as a "vote." The majority of the models' forecasts serve as the basis for the final projection. | Depending on how well each classifier performs, different weights are given to the final prediction. | To improve accuracy, the predictive power of each classifier is restricted by the learning rate. |
| 4. | Each of the subsets is fitted with a decision tree model.  The average of all the decision tree forecasts is used to determine the final prediction. | For instance, if you asked five of your coworkers to give your movie a rating (out of 5), we'll suppose that three of them gave it a 4, and two gave it a 5. Since the majority of respondents rated it a 4, that will be the final rating. | It assigns weights to both classifiers and observations to maximize data variance. | It creates trees based on the residuals of the previous classifier, thereby capturing data variance. |

## **3.5 Experiments**

Extensive experiments were run to measure the performance of the classifiers (Decision Tree DT, K-Nearest Neighbor KNN, Logistics Regression LR, Bernoulli Naïve Bayes NB, Neural Network NN, Quadratic Discriminant Analysis QDA, and Support vector machine SVM) used during the project. Furthermore, extensive experiments were also run to measure the performance of our Ensemble learning models (Ada Boosting, Gradient Boosting, Random Forest, and Max Voting) in detecting Virtual harassment.

## **3.6 Performance Metrics**

To evaluate the performance of the traditional classifiers and ensemble learning on the test data where the true values are known a confusion matrix was used. The performance measures considered in this project include accuracy, recall, precision, and the f1 measure which is determined from the confusion matrix.

* True Positive(TP): This instance indicates Virtual harassment that was classified as Virtual harassment
* True Negative (TN): This instance indicates non-Virtual harassment samples that were classified as non-Virtual harassment.
* False Positive (FP): This instance indicates Virtual harassment samples that were classified as non-Virtual harassment.
* False Negative: It indicates non-Virtual harassment samples that were classified as Virtual harassment.Performance metrics such as accuracy, sensitivity, and specificity are the most widely used in medicine and biology.

The performance metrics are presented mathematically below:

……………………………………………....(3.21)

……………………...…………………….……(3.22)

…………………………..………….……....(3.23)

……………………………………………….……..……….(3.24)

The labeled texts are vectorized which helps to extract features or tokens from each text and a numerical value is assigned to each word or token or feature in the text. Then test datasets are generated using the 70percent of the dataset available used to train the classifiers while 30 percent is used to test the classifiers and the output of each classifier is displayed in a confusion matrix where you have True Positive (Labeled as positive and predicted as positive), True Negative (labeled as negative and predicted as negative ), False Positive (labeled as negative but predicted as positive) and False Negative (labeled as positive but predicted as negative) and performance parameters such as Accuracy, Precision, Recall, F-measure, specificity, MCC, KAPPA, AUC- Area Under Curve, FDR-False Discovery Rate, FNR-False Negative Rate, FPR-False Positive Rate, and NPV-Negative Predictive Value are calculated from the confusion matrix.

### **Table 3.3: THE CONFUSION MATRIX**

|  |  |  |
| --- | --- | --- |
|  | **PREDICTED POSITIVE** | **PREDICTED NEGATIVE** |
| **ACTUAL POSITIVE** | TRUE POSITIVE (TP) | FALSE NEGATIVE (FN) |
| **ACTUAL NEGATIVE** | TRUE NEGATIVE (TN) | FALSE POSITIVE (FP) |

The results are also represented utilizing a boxplot. They enable us to study the distributional characteristics of a group of results. Some important terms are described below:

* **MEDIAN** (middle quartile): denotes the mid-point of the data and is shown by the line that divides the box into two parts. It represents the most likely score.
* **INTER-QUARTILE RANGE**: The middle “box” denotes the central fifty percent of scores for the group. The range of scores from the lower to upper quartile is referred to as the interquartile range. Half of all scores will usually fall in this range.
* **UPPER QUARTILE:** The upper quarter of all scores. Seventy-five percent of the scores fall below the upper quartile.
* **LOWER QUARTILE:** The lower quarter of all scores. Twenty-five percent of scores fall below the lower quartile.
* **WHISKERS:** The upper and lower whiskers represent scores outside the middle 50%. Whiskers often (but not always) stretch over a wider range of scores than the middle quartile groups.

SOCIAL NETWORK API

DATA STREMING

Select Input Data

Read Input Data

Success

No

Yes

VECTORIZATION

TRAINGING THE ALGORITHMS (DECISION TREES, K-NEAREST NEIGHBOR, LOGISTICS REGRESSION, NAIVE BAYES, NEURAL NETWORK, QUADRATIC DISCRIMINANT ANALYSIS, SUPPORT VECTOR MACHINE)

CLASSIFICATION

RESULT (PERFORMANCE EVALUATION USING CONFUSION MATRIX) AND ENSEMBLE LEARNING METHODS (ADA-BOOST, GRADIENT-BOOST, RANDOM FOREST AND MAX VOTING)

### **Figure3.2: Flow Chart For The Text Classification**

Social Networks

Dataset Repository

Based on cyberbullying

Based on non-cyberbullying

Generate training dataset for texts

Texts pre-processing

Textblob

Vectorization

Non-harassment

harassment

Feature extraction

Training dataset

Test Dataset

Classifiers

Non-harassment

Harassment

### **Figure3.3: Process Of Analysis**

# **CHAPTER FOUR**

# **SYSTEM IMPLEMENTATION AND TESTING**

## **Data Analysis for Dataset 1**

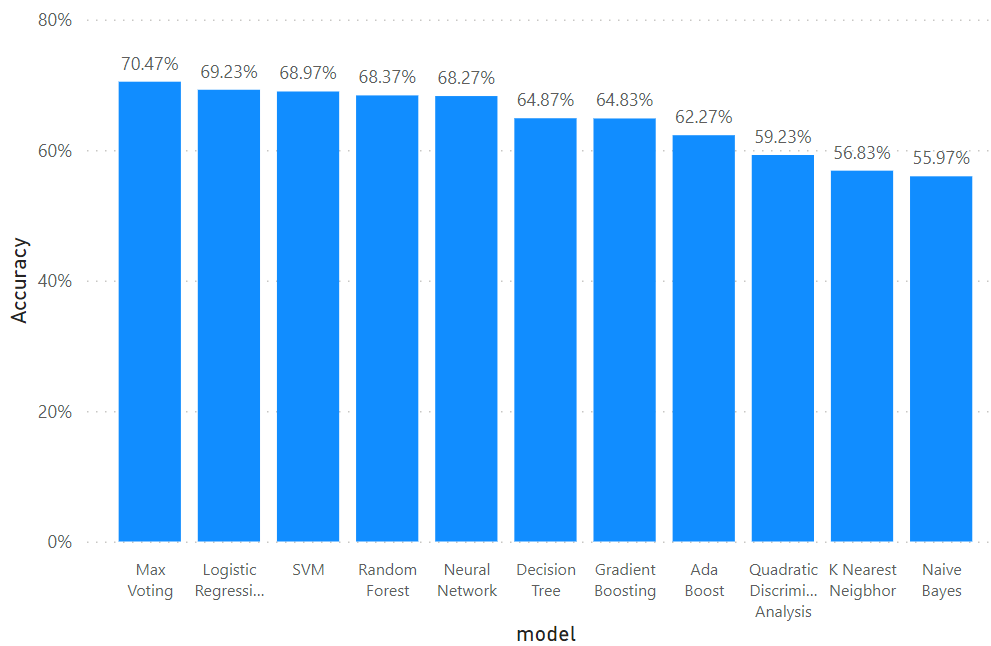
The seven classifiers used in the course of this project include Decision Trees, K-nearest Neighbor, Logistics Regression, Naïve Bayes, Neural Network, Quadratic Discriminant Analysis, Support vector machine (SVM) as well as four Ensemble learnings which are: Ada Boost, Gradient Boosting, Random Forest algorithms, and Max voting were extensively trained using our dataset and outputs were gotten from each of the Algorithms. Each of the algorithms considering our performance metrics, as stated in chapter three, produced an output result consisting of Accuracy, Precision, Recall, Specificity, MCC, KAPPA, F1 Score, AUC, FDR, FNR, FPR, and NPV. The accuracy levels help us know which of the classifiers performs best when it comes to detecting virtual harassment.

Table 4.1 shows the result of the machine learning Algorithm and ensemble learning for Dataset 1 (kaggle, 2021), Logistics Regression has the highest accuracy score of 0.6923, followed by the Support Vector Machine and Neural Network which both attained an accuracy of 0.6897 and 0.6827 respectively, Decision Trees followed with an accuracy of 0.6487, then K Nearest Neighbor, KNN and Quadratic Discriminant Analysis, QDA had an accuracy level of 0.5683 and 0.5923 respectively which is also very encouraging. Gaussian NB had an accuracy level of 0.5597 a very low accuracy result of the seven machine algorithms classifiers used. Figure 4.1 gives the graphical representation of all 7 machine learning algorithms when plotted against the performance metrics.

Figure 4.1 gives a clear graphical representation of the Evaluation metrics and the Algorithms using Dataset 1. The Bar graph represents each of the machine learning algorithms and the results they produced across each of the Evaluation metrics used.

### **Table 4.1** **Result of Dataset 1 (kaggle, 2021) using the Machine Learning Algorithms and Ensemble learning methods**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **specificity** | **MCC** | **KAPPA** | **F1 Score** | **AUC** | **FDR** | **FNR** | **FPR** | **NVP** |
| Decision Trees | 0.6487 | 0.6500 | 0.6500 | 0.6500 | 0.3000 | 0.3000 | 0.6500 | 0.6500 | 0.3500 | 0.3500 | 0.3500 | 0.6500 |
| K-Nearest Neighbor | 0.5683 | 0.5700 | 0.5700 | 0.5700 | 0.1900 | 0.1200 | 0.5700 | 0.5600 | 0.4300 | 0.4300 | 0.4300 | 0.5700 |
| Logistic  regression | **0.6923** | 0.6900 | 0.6900 | 0.6900 | 0.3900 | 0.3900 | 0.6900 | 0.6900 | 0.3100 | 0.3100 | 0.3100 | 0.6900 |
| Gaussian Naïve Bayes | 0.5597 | 0.5600 | 0.5600 | 0.5600 | 0.0800 | 0.0600 | 0.5600 | 0.5400 | 0.4400 | 0.4400 | 0.4400 | 0.5600 |
| Neural Network | 0.6827 | 0.6800 | 0.6800 | 0.6800 | 0.3600 | 0.3600 | 0.6800 | 0.6800 | 0.3200 | 0.3200 | 0.3200 | 0.6800 |
| Quadratic Discriminant Analysis | 0.5923 | 0.5900 | 0.5900 | 0.5900 | 0.2100 | 0.1900 | 0.5900 | 0.6000 | 0.4100 | 0.4100 | 0.4100 | 0.5900 |
| SVM | 0.6897 | 0.6900 | 0.6900 | 0.6900 | 0.3800 | 0.3800 | 0.6900 | 0.6900 | 0.3100 | 0.3100 | 0.3100 | 0;.6900 |
| AdaBoost | 0.6227 | 0.6200 | 0.6200 | 0.6200 | 0.2900 | 0.2500 | 0.6200 | 0.6300 | 0.3800 | 0.3800 | 0.3800 | 0.6200 |
| Gradient Boosting | 0.6483 | 0.6500 | 0.6500 | 0.6500 | 0.3200 | 0.3000 | 0.6500 | 0.6500 | 0.3500 | 0.3500 | 0.3500 | 0.6500 |
| Random Forest | 0.6837 | 0.6800 | 0.6800 | 0.6800 | 0.3700 | 0.3700 | 0.6800 | 0.6800 | 0.3200 | 0.3200 | 0.3200 | 0.6800 |
| Max Voting | **0.7047** | 0.7000 | 0.7000 | 0.7000 | 0.4100 | 0.4100 | 0.7000 | 0.7000 | 0.3000 | 0.3000 | 0.3000 | 0.7000 |



### **Figure 4.1: Graphical representations of Evaluation Metrics against Machine Learning Algorithms and ensemble learning for Dataset 1**

The four ensemble learning models used in the course of this project which include (Ada Boost, Gradient Boosting, Random Forest, and Max Voting) were extensively trained using our datasets, and outputs were gotten from each of the Algorithms.

Each of the algorithms considering our performance metrics, as stated in chapter three, produced an output result consisting of Accuracy, Precision, Recall, Specificity, MCC, KAPPA, F1 Score, AUC, FDR, FNR, FPR, and NPV. The accuracy levels help us know which of the classifiers performs best when it comes to detecting virtual harassment.

Considering the ensemble learning models on dataset 1, The result for the Ensemble classifiers on dataset 1, began with Ada Boost which had the lowest accuracy level at 0.6227, followed by Gradient Boosting and Random Forest which were at 0.6483 and 0.6837 respectively. Overall best ensemble performer for Dataset 1 is the Max Voting ensemble which attained an accuracy of 0.7047 which is still the best in detecting a virtual harassment post in dataset 1, although it produced the lowest FDR, FNR, and FPR at 0.30 (30%). On the other hand, Ada Boost has the lowest MCC and KAPPA of all the four ensemble learning methods at 0.29 and 0.25 respectively which is the sum of errors made for each example during the training and validation process.

Figure 4.3 gives a clear graphical representation of the Evaluation metrics and the Algorithms using Dataset 1. The Bar graph represents each of the ensemble learning models and the results they produced across each of the Evaluation metrics used.

### **Figure 4.2: Graphical representation of Evaluation Metrics against Ensemble Learning Models for Dataset 1**

## **4.1 Data Analysis for Dataset 2**

Table 4.2 Using dataset 2 gives, the result of 7 machine learning algorithms trained, upon successful training of our models using the dataset (YouTube, 2022), results were produced for each of the Algorithms using the evaluation metrics stated earlier. Quadratic Discriminant Analysis QDA had the worst score of 0.1298, the Random Forest ensemble had the highest accuracy score of 0.8779, and KNN, Logistics Regression, and SVM all had accuracy scores of 0.8769 each, same applies to Gradient Boosting and Max Voting ensemble which both had an accuracy score of 0.8769 each.

Following immediately is the Neural Network which had an accuracy level of 0.8731, followed by Ada Boost Ensemble which returned an accuracy score of 0.8654 which is still very okay considering the margin of the other three ensemble learnings, though it seems very low compared to the first three mentioned earlier (Random Forest, Max Voting and Gradient Boosting). Decision Tree had an accuracy level of 0.8298 which is also on the average accuracy score. Gaussian NB had the least accuracy level of 0.7981 which makes it suitable for detecting virtual harassment in comparison with the other accuracy score.

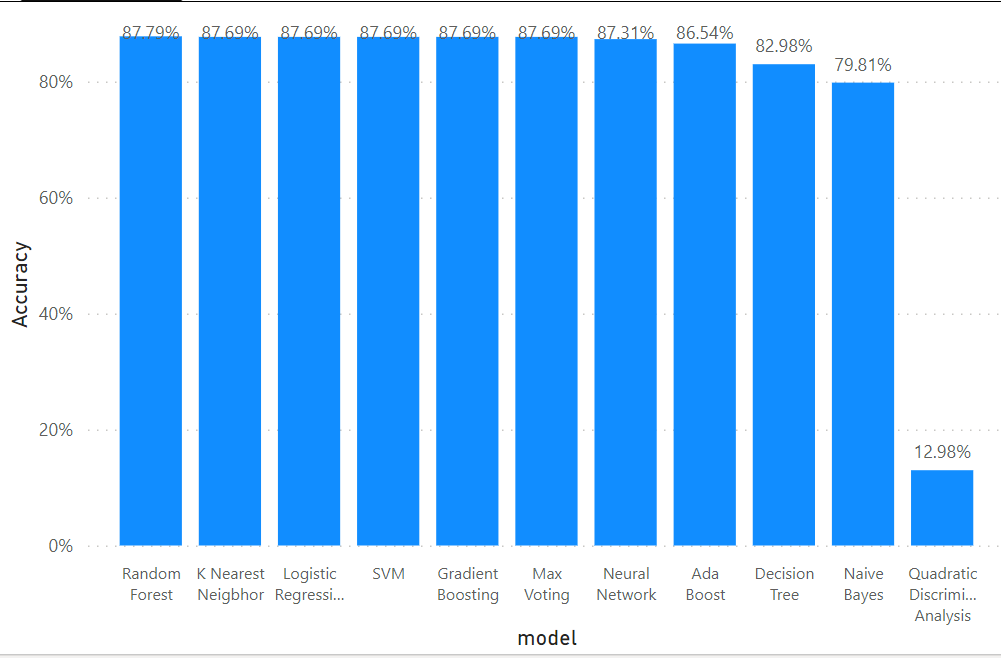
In terms of other performance metrics like MCC and KAPPA, KNN, Logistics Regression, SVM, Max Voting, and QDA all had the worst score of 0.0000, they all returned zero values. The Ada Boost, Gradient Boosting, Random Forest, and Max voting ensembles all performed poorly at FDR, FNR, and FPR respectively as they produced results between 0.12 and 0.13.

Furthermore, as shown in the table, Random Forest which has the highest accuracy also turns out to have a high precision score, of about 0.8779, same applies to its recall, specificity, F1 measure, and Negative Predictive Value (NPV). Random Forest had an average score for the AUC area of 0.50, the same average score applies to all other algorithms and ensemble learning, which all have an average between 0.50 and 0.56.

Figure 4.5 gives a clear graphical representation of the Evaluation metrics, the Algorithms, and ensemble learning techniques using Dataset 2. The Bar graph represents each of the machine learning algorithms and ensemble learning and the accuracy results they produced across.

### **Table 4.2** **Result of Dataset 2** (YouTube, 2022) **using the Machine Learning Algorithms and Ensemble learning methods**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Models** | **Accuracy** | **Precision** | **Recall** | **specificity** | **MCC** | **KAPPA** | **F1 Score** | **AUC** | **FDR** | **FNR** | **FPR** | **NVP** |
| Decision Trees | 0.8298 | 0.8383 | 0.8383 | 0.8383 | 0.1000 | 0.1000 | 0.8300 | 0.5400 | 0.1700 | 0.1700 | 0.1700 | 0.8300 |
| K-Nearest Neighbor | **0.8769** | 0.8800 | 0.8800 | 0.8800 | 0.0000 | 0.0000 | 0.8800 | 0.5000 | 0.1200 | 0.1200 | 0.1200 | 0.8800 |
| Logistic  regression | **0.8769** | 0.8800 | 0.8800 | 0.8800 | 0.0000 | 0.0000 | 0.8800 | 0.5000 | 0.1200 | 0.1200 | 0.1200 | 0.8800 |
| Gaussian Naïve Bayes | 0.7981 | 0.8000 | 0.8000 | 0.8000 | 0.0100 | 0.0100 | 0.8000 | 0.5100 | 0.2000 | 0.2000 | 0.2000 | 0.8000 |
| Neural Network | 0.8731 | 0.8700 | 0.8700 | 0.8700 | 0.1100 | 0.0700 | 0.8700 | 0.5200 | 0.1300 | 0.1300 | 0.1300 | 0.8700 |
| Quadratic Discriminant Analysis | 0.1298 | 0.1300 | 0.1300 | 0.1300 | 0.0300 | 0.0000 | 0.1300 | 0.5000 | 0.8700 | 0.8700 | 0.8700 | 0.1300 |
| SVM | **0.8769** | 0.8800 | 0.8800 | 0.8800 | 0.0000 | 0.0000 | 0.8800 | 0.5000 | 0.1200 | 0.1200 | 0.1200 | 0.8800 |
| AdaBoost | 0.8654 | 0.8700 | 0.8700 | 0.8700 | 0.1800 | 0.1600 | 0.8700 | 0.5600 | 0.1300 | 0.1300 | 0.1300 | 0.8700 |
| Gradient Boosting | 0.8769 | 0.8800 | 0.8800 | 0.8800 | 0.1100 | 0.0600 | 0.8800 | 0.5200 | 0.1200 | 0.1200 | 0.1200 | 0.8800 |
| Random Forest | **0.8779** | 0.8800 | 0.8800 | 0.8800 | 0.0800 | 0.0100 | 0.8800 | 0.5000 | 0.1200 | 0.1200 | 0.1200 | 0.8800 |
| Max Voting | 0.8769 | 0.8800 | 0.8800 | 0.8800 | 0.0000 | 0.0000 | 0.8800 | 0.5000 | 0.1200 | 0.1200 | 0.1200 | 0.8800 |



### **Figure 4.3: Graphical representations of Evaluation Metrics against Machine Learning Algorithms and ensemble learning for Dataset 2**

The four ensemble learning models used in the course of this project which include (Ada Boost, Gradient Boosting, Random Forest, and Max Voting) were extensively trained using our datasets, and outputs were gotten from each of the Algorithms.

Each of the algorithms considering our performance metrics, as stated in chapter three, produced an output result consisting of Accuracy, Precision, Recall, Specificity, MCC, KAPPA, F1 Score, AUC, FDR, FNR, FPR, and NPV. The accuracy levels help us know which of the classifiers performs best when it comes to detecting virtual harassment.

Considering the ensemble learning models on dataset 2, The result for the Ensemble classifiers on dataset 2, beginning with Ada Boost which had the lowest accuracy level at 0.8654, followed by Gradient Boosting and Max Voting which were both a bit higher by 0.0115 each resulting to an accuracy score at 0.8769 for both. Overall best ensemble performer for Dataset 2 is the Random Forest ensemble which attained an accuracy of 0.8779 making it the most appropriate in detecting a virtual harassment post in dataset 2, although it produced poorly in FDR, FNR, and FPR at 0.12 (12%) but not as bad as MCC and KAPPA which both returned 0.00 (0%). On the other hand, Ada Boost has the highest MCC and KAPPA of all the four ensemble learning methods at 0.18 and 0.16 respectively which is the sum of errors made for each example during the training and validation process.

Figure 4.7 gives a clear graphical representation of the Evaluation metrics and the Algorithms using Dataset 2. The Bar graph represents each of the ensemble learning models and the results they produced across each of the Evaluation metrics used.

**Figure 4.4: Graphical representation of Evaluation Metrics against Ensemble Learning Models for Dataset 2**

## **4.2 Data Analysis for Dataset 3**

Table 4.3 gives the result of the 7 Machine algorithms and 4 ensemble learning using dataset 3, Decision Tree attained the accuracy level of 0.9234, KNN had an accuracy of 0.9051, and Logistics Regression also has an accuracy of 0.8896. SVM, QDA, and Neural Network also had accuracy results of 0.9243, 0.8528, and 0.9186 respectively. Gaussian NB on the other hand didn’t do so well in terms of accuracy, having an accuracy level of 0.8070 making it not so far from the true value but still within range.

Table 4.3 also has other Evaluation metrics result; such as Precision, Recall, Specification, F1 Score, and NPV we used across the following algorithms ranging from Decision Trees, KNN, Logistics Regression, Gaussian NB, Neural Network, QDA, and SVM, all attained approximate scores of 0.92, 0.91, 0.89, 0.81, 0.92, 0.85, 0.92 respectively.

The score obtained for FDR, FNR, and FPR across all 7 algorithms range between 0.07 and 0.19 i.e. the range is between 7% …… 19%. This result implies that only one evaluation metric cannot fully be enough to predict the efficiency of an algorithm, except tested with other metrics.

### **Table 4.3** **Result of Dataset 3** (Bayzick, 2022) **using the Machine Learning Algorithms and Ensemble learning methods**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **specificity** | **MCC** | **KAPPA** | **F1 Score** | **AUC** | **FDR** | **FNR** | **FPR** | **NVP** |
| Decision Trees | 0.9234 | 0.9200 | 0.9200 | 0.9200 | 0.7600 | 0.7600 | 0.9200 | 0.8800 | 0.0800 | 0.0800 | 0.0800 | 0.9200 |
| K-Nearest Neighbor | 0.9051 | 0.9100 | 0.9100 | 0.9100 | 0.7000 | 0.7000 | 0.9100 | 0.8400 | 0.0900 | 0.0900 | 0.0900 | 0.9100 |
| Logistic  regression | 0.8896 | 0.8900 | 0.8900 | 0.8900 | 0.6300 | 0.6000 | 0.8900 | 0.7600 | 0.1100 | 0.1100 | 0.1100 | 0.8900 |
| Gaussian Naïve Bayes | 0.8070 | 0.8100 | 0.8100 | 0.8100 | 0.6000 | 0.5500 | 0.8100 | 0.8600 | 0.1900 | 0.1900 | 0.1900 | 0.8100 |
| Neural Network | 0.9186 | 0.9200 | 0.9200 | 0.9200 | 0.7600 | 0.7500 | 0.9200 | 0.8800 | 0.0800 | 0.0800 | 0.0800 | 0.9200 |
| Quadratic Discriminant Analysis | 0.8528 | 0.8500 | 0.8500 | 0.8500 | 0.6700 | 0.6400 | 0.8500 | 0.8900 | 0.1500 | 0.1500 | 0.1500 | 0.8500 |
| SVM | **0.9243** | 0.9200 | 0.9200 | 0.9200 | 0.7600 | 0.7600 | 0.9200 | 0.8700 | 0.0800 | 0.0800 | 0.0800 | 0.9200 |
| AdaBoost | 0.8238 | 0.8200 | 0.8200 | 0.8200 | 0.3500 | 0.3200 | 0.8200 | 0.6300 | 0.1800 | 0.1800 | 0.1800 | 0.8200 |
| Gradient Boosting | 0.8266 | 0.8300 | 0.8300 | 0.8300 | 0.3500 | 0.2300 | 0.8300 | 0.5800 | 0.1700 | 0.1700 | 0.1700 | 0.8300 |
| Random Forest | **0.9258** | 0.9300 | 0.9300 | 0.9300 | 0.7700 | 0.7700 | 0.9300 | 0.8800 | 0.0700 | 0.0700 | 0.0700 | 0.9300 |
| Max Voting | 0.9254 | 0.9300 | 0.9300 | 0.9300 | 0.7700 | 0.7700 | 0.9300 | 0.8700 | 0.0700 | 0.0700 | 0.0700 | 0.9300 |

Figure 4.9 gives a clear graphical representation of the Evaluation metrics and the Algorithms using Dataset 3. The Bar graph represents each of the machine learning algorithms and the accuracy results they produced in percentages.



### **Figure 4.5: Graphical representations of Evaluation Metrics against Machine Learning Algorithms and ensemble learning for Dataset 3**

The four ensemble learning models used in the course of this project which include (Ada boost, Gradient Boosting, Random Forrest, and Max voting as mentioned earlier) were extensively trained using our datasets, and outputs were gotten from each of the Algorithms. Each of the algorithms considering our performance metrics, as stated in chapter three, produced an output result consisting of Accuracy, Precision, Recall, Specificity, MCC, KAPPA, F1 Score, AUC, FDR, FNR, FPR, and NPV. The accuracy levels help us know which of the ensemble classifiers performs best when it comes to detecting virtual harassment

Table 4.4 also gives the result of the 4 ensemble learning models highlighted in green color. Random Forest had the highest accuracy level of 0.9258, Max Voting had an accuracy level of 0.9254, Gradient Boosting had 0.8266 and Ada Boost had the least accuracy level of 0.8238. While considering other evaluation metrics, Random Forest turns out to have a KAPPA score of 0.77(77%) while Gradient Boosting had 0.23 (23%). All four ensemble learning obtained Precision, recall, specificity, F1 score, and NPV of 82%, 83%, 93%, and 93% respectively, starting from Ada Boost, Gradient Boosting, random forest, and Max voting, making it very suitable for detecting virtual harassment in dataset 3, QDA produced an AUC score of 0.89 making it the highest under AUC, showing the perfection in the model’s prediction.

Figure 4.11 gives a clear graphical representation of the Evaluation metrics and the Algorithms and the ensemble learning using Dataset 3. The Bar graph represents each of the ensemble learning models and the results they produced across each of the Evaluation metrics used.

### **Figure 4.6: Graphical representation of Evaluation Metrics against Ensemble Learning Models for Dataset 3**

## **4.3 Data Analysis for Dataset 4**

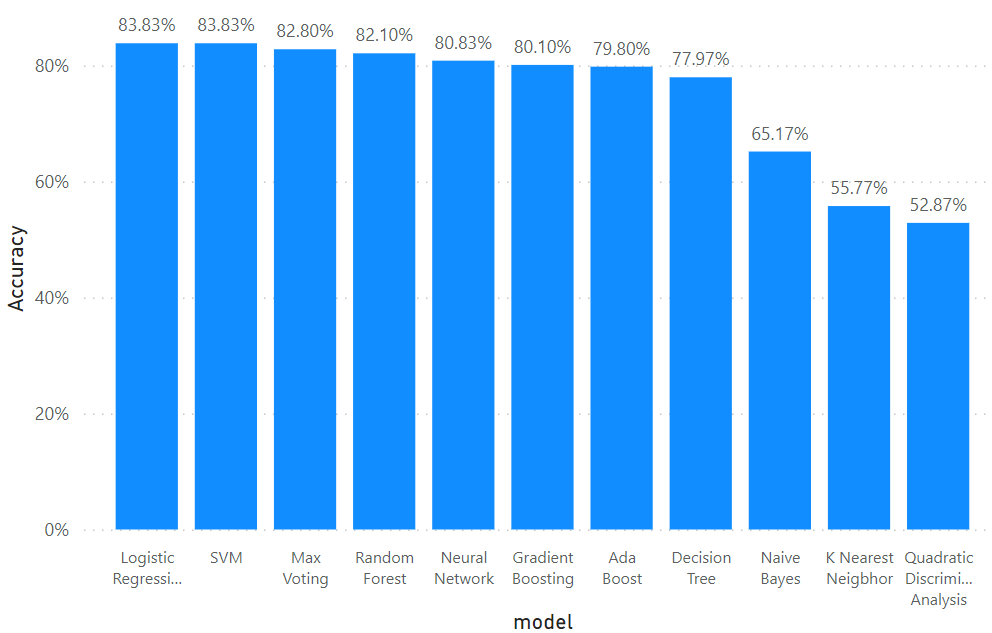
Table 4.4 gives the result of the 7 Machine algorithms as well as their ensemble learning counterpart using dataset 4, Logistics Regression and SVM turns out to have the same highest accuracy level of 0.8383 respectively, Neural Network has an accuracy of 0.8083 and Decision Tree had accuracy results of 0.7797, following immediately is Gaussian NB with an accuracy of 0.6517, KNN and QDA also had accuracy results of 0.5577 and 0.5287 respectively which is on the average level. But the best case matrices for dataset 4 are Accuracy, Precision, recall, specificity, F1 score, and AUC. The second-best performance matrices are MCC and KAPPA, followed by the worst case which are FDR, FNR, and FPR.

Table 4.4 also has other Evaluation metrics results; QDA on the other hand didn’t do so well across the 12 performance matrices. We can see clearly from table 4.4, that the QDA obtained an accuracy of 0.5287, precision, recall, specificity, F1 score, and NPV of 0.53 (53%) respectively, while FNR, FDR, and FPR all attained a score of 0.47 (47%), AUC had the best score of 0.54 (54%), MCC and KAPPA obtained the worst score at 0.09 (9%) and 0.07 (7%) respectively. This result implies that only one evaluation metric cannot fully be enough to predict the efficiency of an algorithm, except tested with other metrics.

Figure 4.13 gives a clear graphical representation of the Evaluation metrics and the Algorithms and ensemble learning using Dataset 4. The Bar graph represents each of the machine learning algorithms and the ensemble learning techniques and the results they produced.

### **Table 4.4** **Result of Dataset 4** (Aggression, 2022) **using the Machine Learning Algorithms and Ensemble learning methods**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **specificity** | **MCC** | **KAPPA** | **F1 Score** | **AUC** | **FDR** | **FNR** | **FPR** | **NVP** |
| Decision Trees | 0.7797 | 0.7800 | 0.7800 | 0.7800 | 0.5600 | 0.5600 | 0.7800 | 0.7800 | 0.2200 | 0.2200 | 0.2200 | 0.7800 |
| K-Nearest Neighbor | 0.5577 | 0.5600 | 0.5600 | 0.5600 | 0.2100 | 0.1000 | 0.5600 | 0.5500 | 0.4400 | 0.4400 | 0.4400 | 0.5600 |
| Logistic  regression | **0.8383** | 0.8400 | 0.8400 | 0.8400 | 0.6800 | 0.6800 | 0.8400 | 0.8400 | 0.1600 | 0.1600 | 0.1600 | 0.8400 |
| Gaussian Naïve Bayes | 0.6517 | 0.6500 | 0.6500 | 0.6500 | 0.3100 | 0.3100 | 0.6500 | 0.6500 | 0.3500 | 0.3500 | 0.3500 | 0.6500 |
| Neural Network | 0.8083 | 0.8100 | 0.8100 | 0.8100 | 0.6200 | 0.6200 | 0.8100 | 0.8100 | 0.1900 | 0.1900 | 0.1900 | 0.8100 |
| Quadratic Discriminant Analysis | 0.5287 | 0.5300 | 0.5300 | 0.5300 | 0.0900 | 0.0700 | 0.5300 | 0.5400 | 0.4700 | 0.4700 | 0.4700 | 0.5300 |
| SVM | **0.8383** | 0.8400 | 0.8400 | 0.8400 | 0.6800 | 0.6800 | 0.8400 | 0.8400 | 0.1600 | 0.1600 | 0.1600 | 0.8400 |
| AdaBoost | 0.7980 | 0.8000 | 0.8000 | 0.8000 | 0.6100 | 0.5900 | 0.8000 | 0.8000 | 0.2000 | 0.2000 | 0.2000 | 0.8000 |
| Gradient Boosting | 0.8010 | 0.8000 | 0.8000 | 0.8000 | 0.6200 | 0.6000 | 0.8000 | 0.8000 | 0.2000 | 0.2000 | 0.2000 | 0.8000 |
| Random Forest | 0.8210 | 0.8200 | 0.8200 | 0.8200 | 0.6500 | 0.6400 | 0.8200 | 0.8200 | 0.1800 | 0.1800 | 0.1800 | 0.8200 |
| Max Voting | **0.8280** | 0.8300 | 0.8300 | 0.8300 | 0.6600 | 0.6500 | 0.8300 | 0.8300 | 0.1700 | 0.1700 | 0.1700 | 0.8300 |



### **Figure 4.7: Graphical representations of Evaluation Metrics against Machine Learning Algorithms and ensemble learning for Dataset 4**

The four ensemble learning models—Ada boost, Gradient Boosting, Random Forest, and Max Voting—used in this study were thoroughly trained using our datasets, and results were obtained from each of the algorithms. Each algorithm that took into account the chapter three performance measures generated an output result that included Accuracy, Precision, Recall, Specificity, MCC, KAPPA, F1 Score, AUC, FDR, FNR, FPR, and NPV. The accuracy levels allow us to identify the ensemble classifiers that perform the best at identifying virtual harassment.

Table 4.4 also gives the result of the 4 ensemble learning models highlighted in green color. Max Voting had the highest accuracy level of 0.8280, Random Forest had an accuracy level of 0.8210, Gradient Boosting had 0.8010 while Ada Boost had the least accuracy level of 0.7980. While considering other evaluation metrics, Ada Boost, Gradient Boosting, Random Forest, and Max voting all turned out to have precisions, recall, F1 measure, specificity, MCC, KAPPA, and NPV of above average making it very suitable for detecting virtual harassment in dataset 4, on the other hand, FDR, FNR, and FPR produced its low loss value of about 0.17 to 0.20, showing it’s an imperfection in the model’s prediction.

Figure 4.15 gives a clear graphical representation of the Evaluation metrics and the Algorithms using Dataset 4. The Bar graph represents each of the ensemble learning models and the algorithms and the results they produced.

### **Figure 4.8: Graphical representation of Evaluation Metrics against Ensemble Learning Models for Dataset 4**

## **4.4 CHOICE OF PROGRAMMING LANGUAGE**

Python is an easy-to-learn, powerful programming language. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. Python’s elegant syntax and dynamic typing, together with its interpreted nature, make it an ideal language for scripting and rapid application development in many areas on most platforms.

Python is an open-source language that many programmers choose to use; it’s a full-featured language with many packages available for fake news detection and data science at large. There are robust libraries and services for testing your code and methods. It’s easy to write defensive code, readability counts, and style matters. It’s straightforward to go from prototype to production. I used PowerBI to collate the results, and software for data virtualization and data analysis.

## **4.5 SOFTWARE AND HARDWARE REQUIREMENTS**

To run all the implementation source code, you are required to install ANACONDA Navigator and launch the JUPITER notebook 6.4.5 which is a web-based, interactive computing environment. It edits and runs human-readable documents while describing the data analysis. You can visit python.org for the installation process and licensing information. The hardware requirements are as follows: a laptop with at least Intel core i5 processors or above, 8 GB RAM, 500GB SDD, (Keyboard and Mouse and VDU - Visual Display Unit / Monitor).

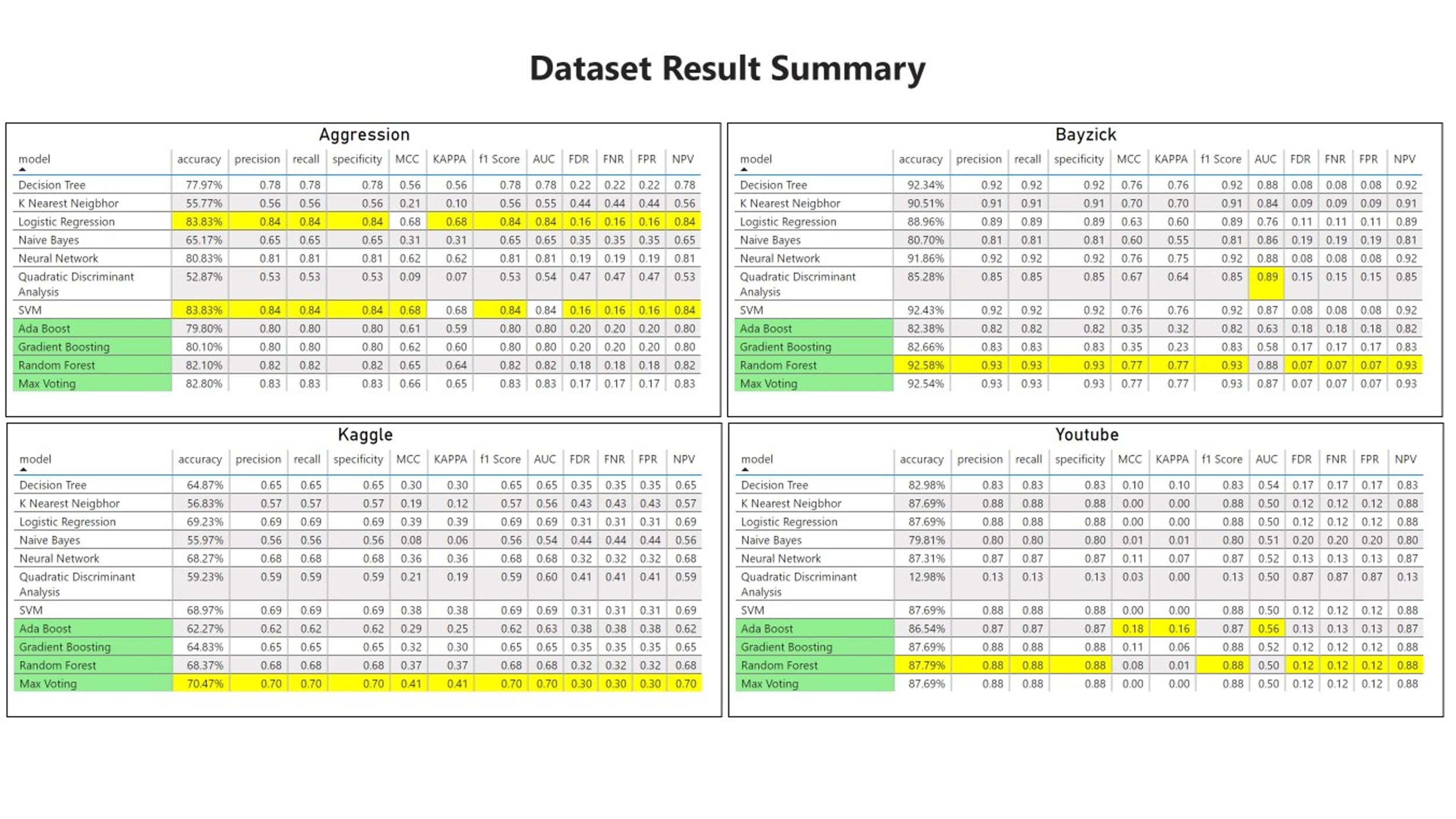


### **Figure4.9 Jupiter Notebook Interface**

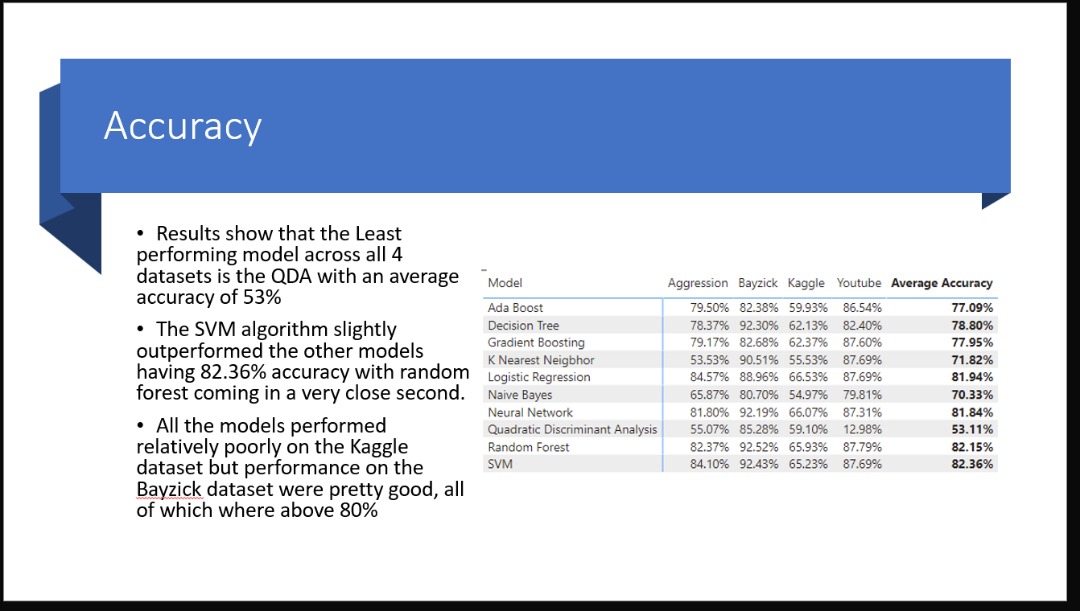
## **4.6 DATA DESCRIPTION**

Datasets from four different platforms were streamed namely: YouTube, Kaggle, Bayzick, and Aggression based on the topic #Harassment and #racism (https://www.kaggle.com/datasets/saurabhshahane/Virtual Harassment-dataset). 30% of the dataset was used to test the algorithms and 70% of the dataset was used to train the algorithm.

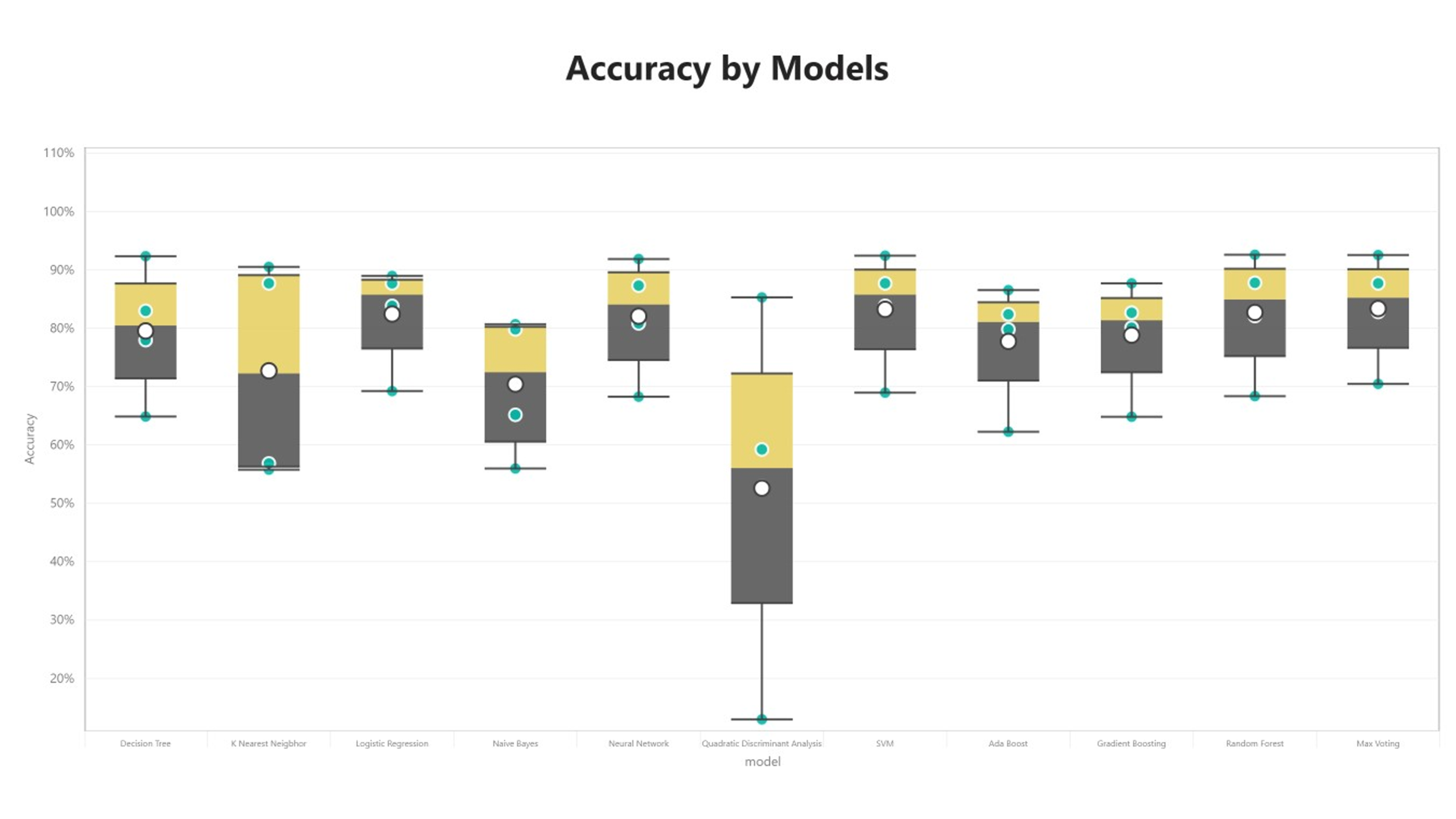
**4.6.1 MODELS VS MATRICES:** The results obtained after implementing eleven algorithms across twelve matrices are shown in Table 4.5. The part labeled green shows the result for the ensemble methods while the yellow part shows the best performing matric in each dataset represented

**Table 4.5: Dataset Result Summary using 11 Model predictions across 12matrices**

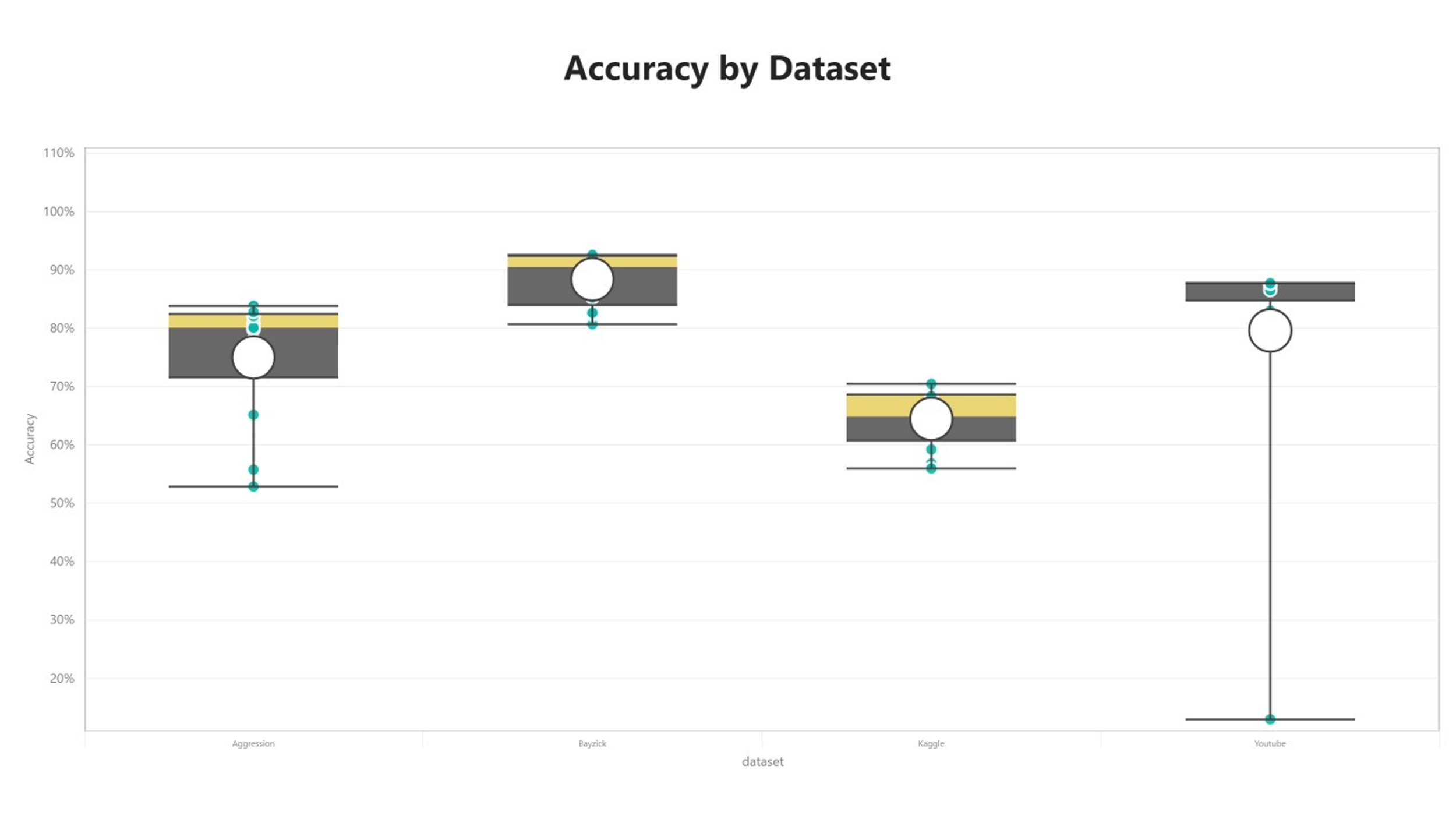
**4.6.2 ACCURACY RESULT:** The results obtained after the implementation of ten different classification algorithms and the Confusion matrix shown in Table 2



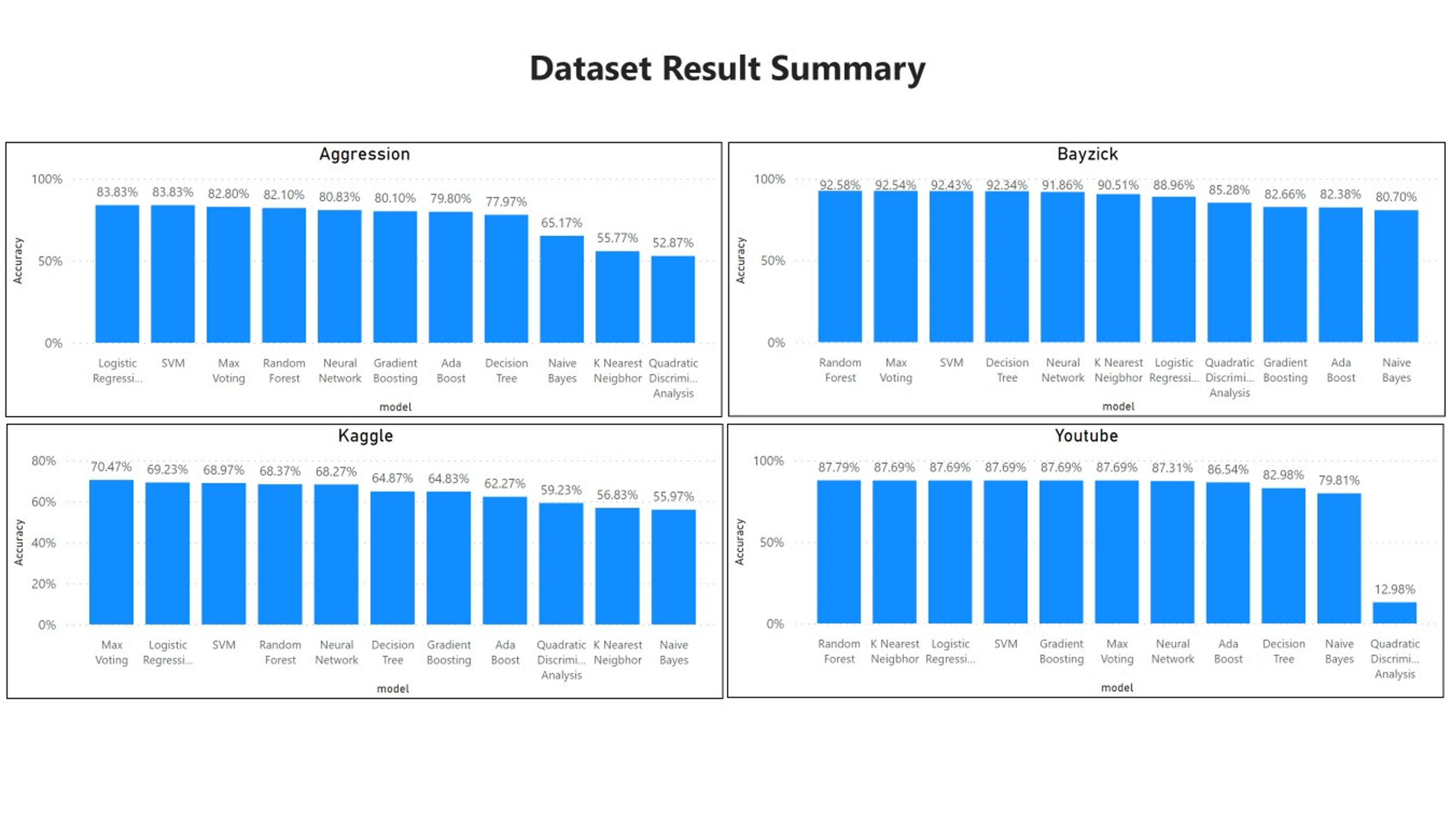
**Table 4.6: Confusion matrix for ten Classification**



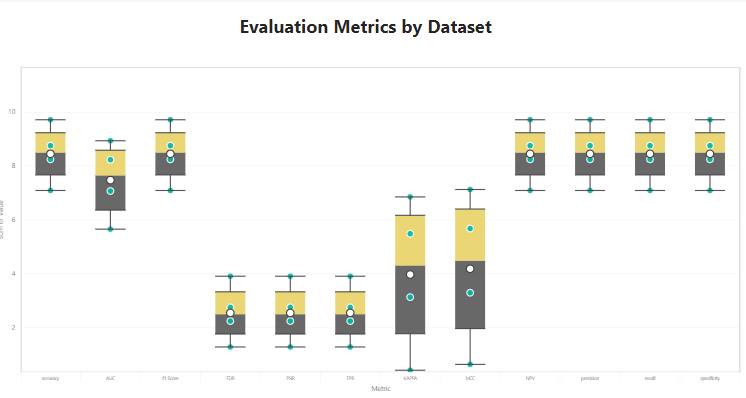
### **Figure4.10 The Comparative Analysis plot based on model Accuracy**



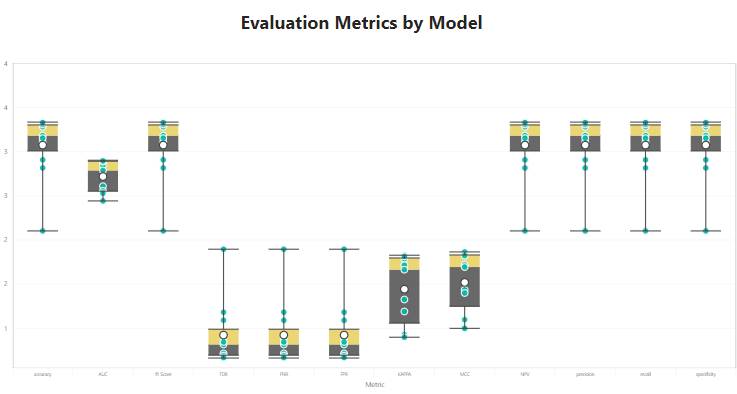
### **Figure 4.11 The Comparative Analysis plot based on the Dataset**



### **Figure 4.12 The Comparative Analysis graph based on models for each dataset.**



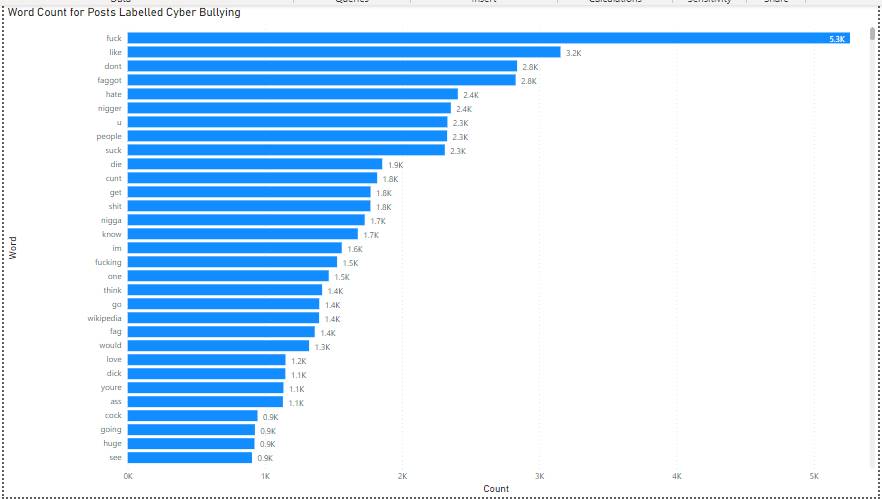
### **Figure 4.13: The evaluation matrices by the four datasets.**

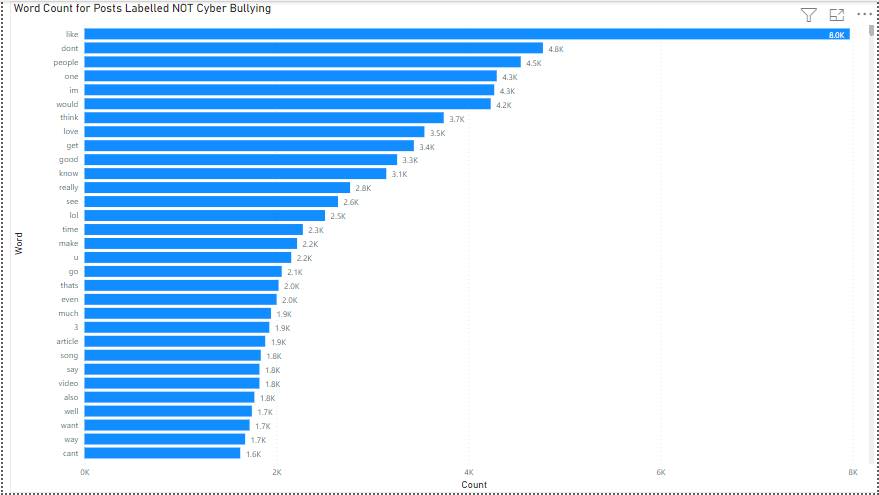


### **Figure4.14 The Plot Average performance of evaluation Metrics by model**

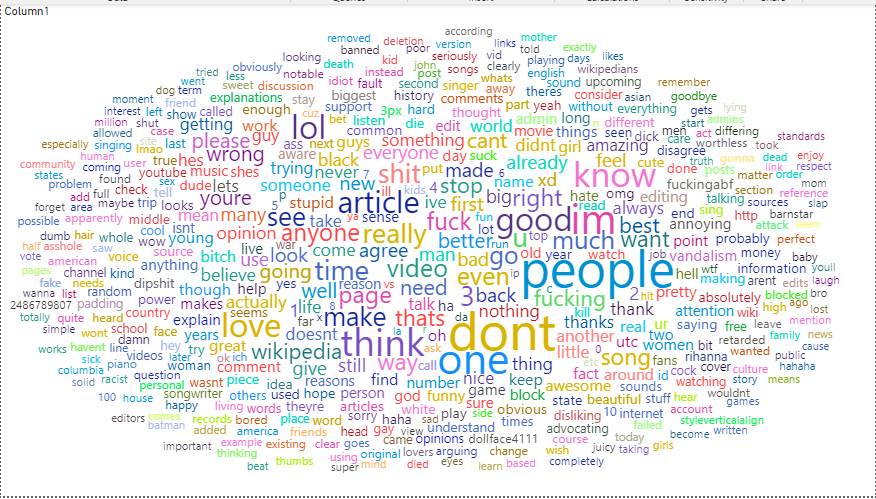
## **4.7 EXPLORATORY DATA ANALYSIS**

This entails the use of keywords in creating a data frame and bar graph of the most common words with their frequency and it was achieved by dropping texts with less than key words by using word split and word count variables:

**Figure4.15 The graph of word count for posts labeled Virtual Harassment and their frequency**

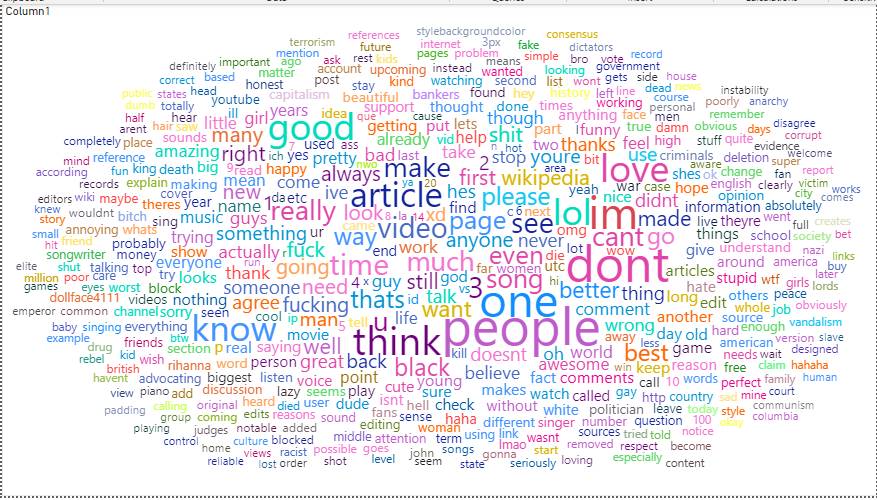


### **Figure4.16 The graph of most common words labeled NOT-Virtual Harassment with frequency**



### **Figure4.17 Word cloud for random unlabeled text**

**Figure4.18 Word cloud for posts labeled to be Virtual Harassment text**



### **Figure4.19 Word cloud for posts labeled non-Virtual Harassment text**

# **CHAPTER FIVE**

# **CONCLUSION AND RECOMMENDATION**

## **5.1 Conclusion**

Virtual harassment is considered to be one of the biggest issues with the development of technology (Internet), social media, and various online communications. It can be carried out by a single user or a group of users who use the internet to harass, embarrass, afflict, torment, and make a nuisance of a specific person online, which has caused serious health problems and is still causing serious health problems to this day, including suicide, depression, and other mental health issues.

To train our machine learning classifiers for classifying comments as Virtual harassment or non-Virtual harassment, a virtual harassment model was developed to detect virtual harassment comments across four different datasets while taking into account the users' features, activity features, and content features. Our machine learning methods, including Decision Tree, K Nearest Neighbor, Logistic Regression, Gaussian NB, Neural Network, Quadratic Discriminant Analysis, and Support Vector Machine, were trained through extensive experiments. Additionally, experiments were conducted utilizing the Ada Boost, Gradient Boosting, Random Forest, and Max Voting ensemble learning models. After utilizing the datasets to train our algorithms, the algorithms were tested and trained using the datasets, the results for the accuracy, precision, recall, specificity, F1 measure, MCC, KAPPA, FDR, FNR, FPR, AUC, and NPV were obtained. Detailed results are shown in Table 4.1–4.7 for the machine learning and ensemble learning models and how they performed across various datasets. Overall, the Bayzick Dataset 3 (Bayzick, n.d.) performed best out of the four datasets used and the worse metrics are the FDR, FNR, and FPR out of the twelve metrics used.

The ensemble learning models outperformed the machine learning models in the evaluation measure produced after the tests because they had access to more data to learn from, as opposed to the machine learning algorithms. When the machine learning algorithms' assessment metrics are contrasted with those for the ensemble learning models, as given in table 4.1-4.7, (in terms of Accuracy, Precision, Recall, Specificity, F1 measure, MCC, KAPPA, FDR, FNR, FPR, AUC, and NPV). Feature engineering is not necessary for ensemble learning models because they are capable of carrying out feature engineering on their own by scanning the dataset for correlated features and combining them for quick learning without being explicitly told to. Although machine learning performs better with small datasets, ensemble learning models need more data to fully realize their potential. As a result, the potential of ensemble learning models is not fully realized with short datasets.

## **5.2 Recommendations**

A user's psychological state at a given time, as well as the study and investigation of seasonal variations in their mood, can have an impact on the language they use to engage in virtual harassment behavior. These changes have an impact on how accurately machine learning algorithms and ensemble learning models can identify virtual harassment. The utilization of long-term data collection on user diversity, psychology, and human behavior is possible. Additionally, because research on virtual harassment in online social media is still in its early stages, results can be further examined using data from other online social media and a repository of already processed datasets.

# **References:**

Abaido, G. M. (n.d.). Virtual Harassment on social media platforms among university students in the United Arab Emirates.

Aggression. (n.d.). *kaggle\_aggressiondataset*. Retrieved from kaggle.com.

Balakrishnan, V. (2015). Cyberbullying among young adults in Malaysia: the roles of gender, age and internet frequency. *Computer in Human Behaviour*(46), 149-157.

Batoul, H., Maroun, C., & Ahmed, S. (2017). A Multilingual System for Cyberbullying detection: Arabic Content detection using machine learning. *Advances in Science, Technology and Engineering System Journal, 2*(6), 275-284.

Bayzick. (n.d.). *Bayzick\_repository*. Retrieved from Bayzick.com.

Cambridge University Press. (2019, December 25). *Cyberbullying*. Retrieved from dictionary.cambridge.org: https://dictionary.cambridge.org/dictionary/english/cyberbullying

Camparitech. (2020). *Cyberbullying facts and statistics for 2020*. Retrieved November 30, 2020, from Camparitech: https://www.comparitech.com/internet-providers/cyberbullying-statistics/

Celestine , I., Gautam, S., Suleman, K., & Praveen , K. M. (2020). Cyberbullying detection based on deep learning architectures.

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting offensive Languages in Social Media to Protect Adolescent Online Safety. *PASSAT and SocialCom*, 22-28.

Chikashi, N., Joel, T., Achint, T., Yashar, M., & YI, C. (2016). Abusive Language detection in online user content. *25th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, 10*(5), 145-153.

Dadvar, M., Trieschnigg, D., & Jong, F. (2014). Experts and Machine against bullies: A hybrid approach to detecting cyberbullies. *A Canadian Artificial Intelligence*(23), 33.

Dictionary. (2019, December 25). *Dictionary.com*. Retrieved from dictionary.com: https://www.dictionary.com/browse/cyberbullying

Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modelling the detection of Textual Cyberbullying. *The Social Mobile Web, 2*(3), 11.

Dipayan Sarkar, V. N. (n.d.). *Ensemble Machine Learning Cookbook.* O'Reilly.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Rasosavljevic, V., & Bhamidipati, N. (2017). Hate Speech Detection with Comment Embeddings. *WWW CyberSafety Workshop*, 12.

Dvořáková, I., & Vochozka, M. (2015). Vykorystannjanejronnychmereždlja prognozuv annj arozytjujompaniji. *Nacional naakedemijamak Ukajainj Kijiv (3)*, 3-5.

Elif, V. A., & Bilal, A. (2018). Detection of Virtual Harassment in Social Networks Using Machine Learning Methods.

GitHub. (2022). *GitHub, Inc*. Retrieved January 25, 2020, from https://github.com/dhfbk/WhatsApp-Dataset

Hine, G., Onaolapo, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Samaras, G., . . . God, T. E. (2017). A measurement Study of 4chain's Politically Incorrectly Forum and its effects on the Web. *WWW CyberSafety Workshop, 3*(2), 11-15.

Hosseinmardi, H., Han, R., Lv, Q., & Mishra, S. (2015). Analyzing Labelled Cyberbullying incidents on the Instagram Social Networks. *IEEE/ACM ASONAM*, 1-7.

Jason, B. (2020, January). *Machine Learning Mastery*. Retrieved September 10, 2020, from Machine Learning Mastery: https://machinelearningmastery.com/category/deep-learning/

kaggle. (2021). *kaggle\_parsed*.

Karthik, D., Birago, J., Catherine, H., Henry, L., & Roslind, P. (2012). Common sense reasoning for detecting, preventing, and mitigating cyberbullying. *ACM Transactions on Interactive Intelligent Systems (IIS), 2*(3), 18.

Kayes, I., Kourtellis, N., Quercia, D., Iamnitchi, A., & Bonchi, F. (2015). The Social World of Content Abusers in Community Questions Answering. *WWW CyberSafety Workshop, 3*(3), 22-24.

Krishna, K. B., & Narendra, S. M. (2015). A framework for Cyberbullying Detection in social media. *International Journal of Current Engineering and Technology, 5*(1), 1-5.

Lenhart, A. (2019, December 23). *Teens, cell phones, and texting.* . Retrieved from Pew Internet & American Life Project: http://pewresearch.org/pubs/1572/teens-cell-phones-text-messages

Lexico. (2019, December 26). *lexico.com*. Retrieved from Lexico: https://www.lexico.com/en/definition/cyberbullying

Liu, W., & Ruths, D. (2013). What's a name? Using the first names as features for gender information on Twitter. *International Conference of Data Mining*, 23-45.

Mahmud, J., Zhou, M., Megiddo, N., Nichola, J., & Drews, C. (2013). Recommending targeted strangers from whom to solicit information on social media. *Proceedings of 2013 International Conference on Intelligent user interface, ACM*, 37-48.

Manuel, F. V., Francisco, N. J., Victor, C., & Fidel, C. (2021). Early Detection of Cyberbullying on Social Media Networks.

Merriam-Webster. (2019, December 25). *Cyberbullying*. Retrieved from Merriam-webster.com: https://www.merriam-webster.com/dictionary/cyberbullying

Michele, D., Emmanuel, D., & Alfredo , P. (2016). Unsupervised Cyberbullying Detection in Social Networks. *23rd International Conference on Pattern Recognition (ICPR)*, 432-437.

Moreno, M., & Kota, R. (2014). *Social Media in B. Strasburger B, Wilson & A. Jordan (Eds.) Children, Adolescents, and the media.* Califonia: Sage Publishers.

Nahar, V., Unankard, S., Li, X., & Pang, C. (2012). Sentiment Analysis for Effective Detection of CyberBullying. *Asia-Pacific Web and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data, III*(12), 22.

Nandhini, B., & Sheeba , J. (2015). Cyberbullying detection and classification using information retrieval algorithm. *In Proceedings of the 2015 International Conference on Advanced Research in Computer Science, Engineering & Technology*, 20.

Nicole, A. M. (2009). Cyberbullying. *The Graduate School, University of Wisconsin-Stout*, 1-33.

Nnamdimi, B., & Sheeba, J. (2015). Cyberbullying detection and classification using Information retrieval Algorithm. *International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET )*, 1-6.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive Language Detection in Online User Content. *World wide Web, IV*(3), 22.

Nureni , A. A., Sunday, I., Chinazo, O. J., & Charles, V. V. (2021). Cyberbullying Detection in Social Networks: Artificial Intelligence Approach. *Journal of Cyber Security and Mobility, Vol. 10 4. doi: 10.13052/jcsm2245-1439.1046*, 745–774.

OrboGraph. (2020). *Orbograph*. Retrieved September 10, 2020, from OrboGraph Website: https://orbograph.com/deep-learning-how-will-it-change-healthcare/

Pieschl, S., Porsh, T., Kahl, T., & Klockenbusch, R. (2013). Relevant dimensions of Cyberbullying -Resuls from two experimental studies. *Jounal of Applied Development Psychology, 5*(34).

Richard, D. (2012). Bullying and Cyberbullying: History, Statistics, Law, Prevention, and Analysis. *The Elon Journal of Undergraduate Research in Communications • Vol. 3, No. 1*, 1-10.

Rui, Z., Anna, Z., & Kezhi, M. (2016). Automatic detection of Cyberbullying on social networks based on bullying features. *Proceedings of the 17th International Conference on distributed computing and Networking, ACM*, 43.

Sani, M., & Livia, A. (2017). Cyberbullying Classification using text mining. *1st International Conference on Information and Computational Science, IEEE*, 241-246.

Saravanaraj, A., Sheeba, J., & Pradeep, D. S. (2016). Automatic Detection of Cyberbullying from Twitter. *Journal of Computer Science and Information Technology, III*(6), 22.

Shane, M., William, B. J., Adrian, S., & Gordon, R. (2018). Machine learning and semantics analysis of in-game chat for Cyberbullying. *Journal on Computers and Security*(76), 197-213.

Shiels, M. (2003, December 23). *A chat with the man behind mobiles. BBC News.* Retrieved from BBC News: http://news.bbc.

Sophia, A. (2016). Cyberbullying in the world of Teenagers and social media: A Literature review. I*nternational Journal of Cyber Behavior, Psychology and Learning Volume 2 Issue 2*, 1-29.

Sourabh, P., & Vaibhav, S. (2014). Cyberbullying detection and prevention: Data mining and psychological perspective. *International Conference on Circuit, Power and Computing Technologies (ICCPCT), 10*(5), 1541-1547.

Sourabh, P., & Vaivhav, S. (2014). Cyberbullying Detection and Prevention: Data Mining psychological perspective. *International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, 10-12.

Springer. (2021). *https://link.springer.com/chapter/10.1007/978-981-15-1967-3\_8*. Retrieved from Ensemble Learning. In: Machine Learning. Springer, Singapore: https://link.springer.com/chapter/10.1007/978-981-15-1967-3\_8

statista. (2022). *Social Networks.*

uslegal. (n.d.). *Cyber Harassment Law and Legal Definition*. Retrieved from https://definitions.uslegal.com/c/cyber-harassment: https://definitions.uslegal.com/c/cyber-harassment

Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., . . . Hoste, V. (2015). Automatic detection and prevention of Cyberbullying. *Human and Social Analytics, II*(2), 12-24.

Walisa, R., Lodchakorn, N. n., Pimpaka, P., Piyaporn, N., & Pirom, K. (2017). Automated Cyberbullying detections using clustering appearance patterns. *9th International Conference on Knowledge and Smart Technology*, 242-247.

Wang, W., Chen, L., Thirnarayan, K., & Sheth, A. (2014). Cursing in English on Twitter. *In Proceedings of the 17th ACM conference on Computer supported cooperative work & Social computing, ACM*, 415-425.

Xiang, Z., Jonathan, T., Nishant, V., Elizabeth, W., Joseph, M. P., Robin, K., . . . Edward, D. (2016). Cyberbullying detection with pronunciation-based Convolutional Neural Network. *15th IEEE International Conference on Machine Learning and Applications (ICMLA), 10*(5), 740-745.

YouTube. (n.d.). *kaggle.com*.