**DETECTION OF CYBERBULLYING WITH CATBOOST CLASSIFICATION AND N-GRAM LANGUAGE MODEL**

**BY**

**ADEBAYO NEHEMIAH PEACE ADEOLUWA**

**(20PGDC000112)**

**DEPARTMENT OF COMPUTER SCIENCE,**

**COLLEGE OF PURE AND APPLIED SCIENCES,**

**LANDMARK UNIVERSITY, OMU-ARAN,**

**KWARA STATE, NIGERIA.**

**AN M.Sc. RESEARCH SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE AWARD OF M.Sc. COMPUTER SCIENCE**

**LANDMARK UNIVERSITY, OMU-ARAN.**

**(MAY, 2022)**

**DECLARATION**

I, Nehemiah Adeoluwa Adebayo, hereby declare that this dissertation I have submitted, titled “**Detection of Cyberbullying with CatBoost Classification and N-Gram Language Model**” is entirely my original work undertaken in the Department of Computer Science, Landmark University, Omu-Aran, and I have properly cited any outside information or work used.

Nehemiah Adeoluwa ADEBAYO: **20PGDC000112**

----------------------------------------

Full Name

----------------------

Signature & Date

**CERTIFICATION**

This is to certify that this dissertation has been read and approved as meeting the requirements of the Department of Computer Science, Landmark University, Omu-Aran, Nigeria, for the Award of Master of Science (M.Sc.) Computer Science.

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

Dr. Amos O. Bajeh Date

(Supervisor)

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

Dr. Moses Abiodun Date

(Co-Supervisor)

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

Dr. M. O. Adebiyi Date

(H. O. D)

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

Prof. R.O. Oladele

(External Examiner) Date

**ABSTRACT**

The exponential rise in the use of social media and the number of users have led to a proportional increase in the frequency of bullying as social media provides a thriving environment for its perpetrators. Instant messaging platforms such as Facebook, Twitter, and WhatsApp are common areas where the effects of online bullying are well pronounced.

The overall negative effects of cyberbullying are not far-fetched as it leads to depression and suicidal tendency among teenage victims. As a result, there is a need to develop measures for detecting cyberbullying, hence, the purpose of this study. Although, in recent years, studies have been conducted to automatically detect instances of cyberbullying on social media with the use of conventional machine learning algorithms. In contrast to other studies, this study is aimed at using the CatBoost classification algorithm in conjunction with the n-gram language model to automatically detect cyberbullying while taking context into account. The language model works by capturing the context, sentiment, and frequency of words from the Twitter dataset. After the training and testing of the CatBoost classification algorithm and the n-gram language model, and other classification algorithm namely Naive Bayes, Support Vector Machine, Random Forest, and Decision Tree, results shows that the CatBoost outperforms other classification algorithm with an accuracy of 97%. In terms of precision and recall, the CatBoost model gave a performance of 90% and 86% respectively when tested with random real-life textual data on the web application framework. The Web application framework developed will aid in detecting instances of cyberbullying in real-life scenarios, thereby reducing the impact of cyberbullying.

**DEDICATION**

This project is dedicated to God Almighty for making this whole programme a success in good health and sound mind. This dedication goes to my Parents Mr and Mrs Adebayo and my siblings for their support and continuous prayer throughout this programme.

**ACKNOWLEDGEMENT**

I want to express my gratitude to the all-powerful and all-sovereign, God, who is the giver of life, knowledge, strength, and other virtues, and who via inspiration conceived of and gave birth to this research endeavour. God has demonstrated and shown His involvement in the play from the beginning to the end.

Bishop David Oyedepo, chancellor of Landmark University: thank you for answering the call of faith and destiny by allowing someone like me the honour of attending "The Amazing Landmark University," as we affectionately called it. Several people worth my acknowledgement for contributing to my accomplishment in this phase of academic studies in one way or the other. But my special appreciation goes to my supervisor, Dr. Amos Orenyi Bajeh, and my Co-supervisor, Dr. Moses Abiodun for their continual mentorship, encouragement and close follow up during my studies.

To the Dean of the College of Pure and Applied Sciences, Prof. Ayodele Adebiyi and H.O.D of the Department of Computer Science, Dr. Mrs. Marion Adebiyi; thanks for the opportunity and acceptance into the department of Computer Science in this amazing University. Landmark University's computer science faculty members have made this programme a memorable one for me, and I would like to thank them all for their efforts in the classroom and on the field. Having Udo, Mary, and Suraj as classmates has been an honour, and I am grateful for the experiences we have made together. The Lord bless you all and reward you according to your good acts to me.

TABLE OF CONTENTS

List of Tables.........................................................................................................................viii

List of Figures........................................................................................................................ix

[**CHAPTER ONE: INTRODUCTION** 1](#_Toc104761837)

[1.1 Background to the Study 1](#_Toc104761838)

[1.2 Statement of the Problem 2](#_Toc104761839)

[1.3 Justification for the Study 3](#_Toc104761840)

[1.4 Aim and Objectives 4](#_Toc104761841)

[1.5 Scope of the Study 5](#_Toc104761842)

[**CHAPTER TWO: REVIEW OF LITERATURE** 6](#_Toc104761843)

[2.1 Cyberbullying 6](#_Toc104761844)

[2.2 Cyberbullying on Social Media 11](#_Toc104761845)

[2.3 Text Classification 13](#_Toc104761846)

[2.4 N-Gram 14](#_Toc104761847)

[2.6 Text Classification Algorithm 16](#_Toc104761848)

[2.7 Survey of Related Works 20](#_Toc104761849)

[2.8 Research Gap 25](#_Toc104761850)

[**CHAPTER THREE: METHODOLOGY** 26](#_Toc104761851)

[3.1 Introduction 26](#_Toc104761852)

[3.3 Research Framework 26](#_Toc104761853)

[3.4 Data Collection 28](#_Toc104761855)

3.5 Balancing the Data 30

[3.6 Text Preprocessing 29](#_Toc104761857)

[3.7 Evaluation Metrics 30](#_Toc104761858)

[**CHAPTER 4: RESULTS AND DISCUSSION** 34](#_Toc104761860)

[4.1 Introduction 34](#_Toc104761861)

[4.3 Datasets 35](#_Toc104761864)

[4.4 Results 37](#_Toc104761867)

[**CHAPTER FIVE: SUMMARY, CONCLUSION AND RECOMMENDATIONS**](#_Toc104761888)

[5.1 Research Summary 50](#_Toc104761889)

5.2 Conclusion 51

[5.3 Recommendation 52](#_Toc104761890)

[**REFERENCES** 53](#_Toc104761891)

**LIST OF TABLES**

[Table 3.1: Details of data collected 2](#_Toc104761856)9

[Table 4.1 Results of SVM model (Mendeley dataset) 3](#_Toc104761868)9

[Table 4.2 Results of Decision Tree model (Mendeley dataset) 3](#_Toc104761869)9

[Table 4.3 Results of Random Forest model (Mendeley dataset) 3](#_Toc104761870)9

[Table 4.4 Results of Naive Bayes model (Mendeley dataset) 4](#_Toc104761871)0

[Table 4.5 Results of CatBoost classifier model (Mendeley dataset) 4](#_Toc104761872)0

[Table 4.6 Results of SVM model (Kaggle dataset) 4](#_Toc104761873)1

[Table 4.7 Results of Decision Tree model (Kaggle dataset) 4](#_Toc104761874)1

[Table 4.8 Results of Random Forest model (Kaggle dataset) 4](#_Toc104761875)2

[Table 4.9 Results of CatBoost classifier (Kaggle dataset) 4](#_Toc104761876)2

[Table 4.10 Results of Naive Bayes (Kaggle dataset) 4](#_Toc104761877)2

[Table 4.11 Results of SVM model (Balanced Mendeley dataset) 4](#_Toc104761878)4

[Table 4.12 Results of Decision Tree model (Balanced Mendeley dataset) 4](#_Toc104761879)4

[Table 4.13 Results of Random Forest model (Balanced Mendeley dataset) 4](#_Toc104761880)5

[Table 4.14 Results of Naive Bayes model (Balanced Mendeley dataset) 4](#_Toc104761881)5

[Table 4.15 Results of CatBoost classifier model (Balanced Mendeley dataset) 4](#_Toc104761882)6

[Table 4.16 Results of SVM model (Balanced Kaggle dataset) 4](#_Toc104761883)7

[Table 4.17 Results of Decision Tree model (Balanced Kaggle dataset) 4](#_Toc104761884)7

[Table 4.18 Results of Random Forest model (Balanced Kaggle dataset) 4](#_Toc104761885)8

[Table 4.19 Results of Naive Bayes model (Balanced Kaggle dataset) 4](#_Toc104761886)8

[Table 4.20 Results of CatBoost classifier model (Balanced Kaggle dataset) 4](#_Toc104761887)9

Table 4.21 Mendeley Imbalanced and Balanced result 49

Table 4.22 Kaggle Imbalanced and Balanced result 50

**LIST OF FIGURES**

Figure 3.1: Research framework 27

Figure 4.1: unigram visualization for text corpus 34

Figure 4.2: bigram visualization for text corpus 34

Figure 4.3: trigram visualization for text corpus 35

Figure 4.4: Mendeley dataset distribution 36

Figure 4.5: Kaggle dataset distribution 36

**CHAPTER ONE**

# INTRODUCTION

## 1.1 Background to the Study

It is easy to lose track of how much time you spend on social media because of how addicting it is and how much data it produces (Van Hee et al., 2018).

The vast majority of users who have access to the internet has increased. As a result, the opportunities for both good and bad behaviour have also increased. One of the risks of using the internet is being subjected to cyberbullying.

Cyberbullying may be defined as bullying that takes place via digital technology, and users of the internet who are subjected to this harassment on online platforms have a higher potential of exhibiting signs of anxiety, body dysmorphia, depression, and even suicidal thoughts. To carry out their acts without fear of being identified, bullies benefit from anonymity. Among the most typical targets of cyberbullying are children and teenagers because of their enthusiasm for advanced technologies such as the internet. Traditional bullying is exceptionally similar to cyberbullying, except that traditional bullying takes place physically, whereas cyberbullying takes place online, where a message can be delivered directly to the victim or put on public media platforms where everyone on that platform can view it. (Alim, 2016)

Regulations designed to combat cyberbullying have been introduced in several countries (Cross et al., 2015), but the results of these efforts are not favorable. As a direct consequence of this, the amount of harmful content, including hate speech and negative material is increasing on social media platforms. According to research (Yang,et al., 2014), cyberbullying has been proven to impact victims' mental health and self-esteem negatively. According to Price and Dalgleish (2010), cyberbullying can hurt student achievements; result in issues with anger control, and even lead to suicide. Based on of all negative factors caused by cyberbullying, its detection and eradication is of paramount importance (Mahesh et al., 2021).

When it comes to retrieving valuable information from unstructured text data, text analytics is a crucial component of the retrieval process. Natural language processing and statistical algorithms perform operations like named entity recognition, sentiment analysis, and text classification; these tasks turn unstructured data into structured text by defining complex patterns and pulling out the most crucial information from them. Text classification is the process of organising unstructured material into a range of categories. The goal of training a text classification model is to reliably predict a label for data that has never been seen before. Text classification can be achieved with the use of a wide variety of models and the type of model to use depends on the data type and how accurate the classification needs to be.

This study makes use of unstructured datasets (tweets) to train a machine learning model in order to help aid the identification cyberbullying text on social media. Natural language processing (NLP) and machine learning can be used to categorise big datasets more effectively. Using a trained model, text classifiers generate an output that indicates whether the review has a positive, negative, or neutral sentiment.

## 1.2 Statement of the Problem

A person's character and personality are often reflected in how they communicate via language. Because of this, various harassers will adopt different writing styles to get their message through.(Arif, 2021). However, regardless of how a word is defined, the semantic categories to which it belongs remain constant. Therefore, rather than focusing on specific phrases, the authors of this study analyze the order and structure of words to spot signs of cyberbullying.

Around the globe, with a wide age range, teenagers to adults have been significantly impacted by the issue of cyberbullying. The current state of the art in detecting cyberbullying is rapidly evolving, but several challenges have emerged (El Asam & Samara, 2016). Most detection algorithms consider offensive words to be features, but they change over time; hence, automated approaches rather than constant features must be used in detection. (Mahesh et al., 2021).

Previous research has encountered difficulties with imbalance classification, as well as overfitting and underfitting issues. Because of these issues, the accuracy of a classification model cannot be relied on because it will be less effective in detecting actual instances of cyberbullying. The primary purpose is to devise a detection method that can take into account semantics and meaning, provide accurate findings, and do all of this while keeping the amount of time and money spent on computation to a minimum. Text and sentimental analytics were heavily used in developing the text classification algorithm that was utilised in this investigation. This was done to carefully analyse textual data and fully grasp the purpose behind each individual word.

## 1.3 Justification for the Study

As the Internet usage rate increases, so does the detrimental effect of cyberbullying. Various investigative and predicting algorithms discussed and developed in recent literature have limits. A comprehensive system that can automatically recognise cyberbullying text is essential, and this study is a step in the right direction to achieve that goal.

Many studies have considered text classification algorithms to find instances of cyberbullying. Still, few have looked at how the CatBoost classifier method may be used to do the same. (Rezvani et al., 2020). Commonly used classification methods have been extensively employed in previous studies; however, classification imbalance, training duration, overfitting and underfitting problems remain unsolved. Therefore, this study provides new insights into the characteristics of the CatBoost classifier and its applications for detecting instances of cyberbullying.

CatBoost efficiently and automatically handles heterogeneous, sparse, and categorical data. When it comes to machine learning implementations for large amounts of data, Hancock and Khoshgoftaa. (2020) findings show CatBoost to be a viable option, especially for datasets that include both categorical and non-categorical features. They opined that CatBoost will continue to be a good choice for many applications for some time because its ease of use and ability to automatically handle categorical data. It also performs better compared to other gradient boosting decision tree implementations. Also, the n-gram text classification language model is used with the CatBoost algorithm to make it even better at recognising cyberbullying from textual data.

## 1.4 Aim and Objectives

This research aims to investigate the performance of CatBoost classification algorithm with N-gram language model for cyberbullying detection.

To achieve this aim, the following objectives are set:

i. to design CatBoost and N-gram language model,

ii. to implement the designed model

iii. to comparatively evaluate the performance of all models with other ML models: Support vector machine, Naive Bayes, Decision tree, and Random forest.

## 1.5 Scope of the Study

This study investigates the functionalities of the CatBoost classifier, with help from a collection of labelled cyberbullying datasets from the Kaggle website and by combining the text classification method with the n-gram language model, we were able to address some of the major problems with text classification encountered in previous studies. After the training and testing phases are completed, the dataset will be trained into an machine learning model, which will then be used to create a web application using Python web framework.

**CHAPTER TWO**

# REVIEW OF LITERATURE

## 2.1 Cyberbullying

The term cyberbullying refers to the act of intimidating, harassing, or threatening another person by using a digital platform. It has been shown that this kind of harassment is associated with a vast number of significant problems relating to mental health. Harmful bullying behaviour includes: Publishing rumors about someone else, making threats, and revealing personal information about the victim.

Repeated action with the intention to do damage is two of the defining characteristics of bullying and harassment. This practice is considered to be a form of online harassment. It has been seen more often, particularly among teenagers in recent years(Pricilla et al., 2019). Cyberbullying may take place on social media platforms anywhere and at any given time, in contrast to traditional forms of bullying, which take place face-to-face. Bullies may inflict misery on their peers as long as they can disguise themselves behind the internet and avoid confrontation. Because victims are continually hooked up to the internet and online social networks, particularly among young people, they are susceptible to being harassed (Mahesh et al., 2021).

**2.1.1 Characteristics of Cyberbullying**

The following are the important aspects of cyberbullying:

1. Persistence: Even after being confronted or punished, some bullies continue to be involved in cyberbullying. Bullies that engage in cyberbullying may be doing so for no other reason than they like inflicting harm on others and believe that there are no repercussions for their actions. The Bullies may also be venting their frustrations on others to distract themselves from their difficulties or insecurities.
2. Anonymity: Since cyberbullying can be anonymous and difficult to trace, it can be difficult for victims to defend themselves. It can also have a lasting impact, as messages and images can be spread quickly and widely anonymously. As a result, it is crucial to be aware of the signs of cyberbullying and to take steps to prevent it. Additionally, the Bullies may feel that they have a sense of power or control when they engage in cyberbullying and can get away with it because they are anonymous.

Cyberbullying is prevalent nowadays as it can take many forms, which are listed below.

i. Trolling is a kind of cyberbullying in which the perpetrator purposefully seeks to infuriate or irritate the victim. Provocative or offensive remarks or messages may be sent, and violent or abusive information may be posted on social media to harass, threaten or intimidate an individual.

ii. Harassment: It covers a variety of bullying actions, but it most commonly refers to a habit of sending harmful or threatening statements to users. A quick step a victim can take to try and stop it is to try and figure out who is behind the harassment; after identifying the person or people responsible, proceed to block them from being able to contact you further should be taken. This may involve changing privacy settings on social media platforms or even blocking the Bully entirely.

i. Doxing: a form of online harassment that involves revealing someone's personal information without permission. In some cases, doxers will even go as far as sharing sensitive information like someone's medical records or financial information. Doxing is often used to intimidate or humiliate someone, and it can have serious real-world consequences. For example, if someone's home address is revealed, they may be at risk of being physically harmed by criminals.

ii. Masquerading: is a form of cyberbullying in which someone creates a fake online profile to harm or harass others. This can be done by impersonating someone else or creating a false identity (CyberSecurity Training n.d.). The goal is to deceive others into damaging their reputation or causing them emotional distress. Masquerading can be incredibly harmful, especially if the victims do not realize they are being deceived. This can lead them to make decisions based on false information or believing things about themselves that are not true.

iii. Dissing: it is the act of sending someone nasty or disrespectful messages. Text messaging, social networking, and video messaging applications are all viable options where it is carried out on victims. It is also possible to be disrespectful by spreading rumours or making derogatory comments about another person on the internet. Disrespect is the root of the word "diss." When someone disses an individual, he/she shows a lack of respect for that individual.

iv. Cyberstalking: As the name suggests, cyberstalking is a kind of cyberbullying in which the stalker targets their victim via internet communications. Sending threatening or vulgar messages, publishing embarrassing photos or information about the victim online, or even hacking into their accounts to steal personal information are examples (Al-Rahmi et al., 2020). In addition, cyberstalkers often use social media to monitor their victims and communicate with them, making it hard for their victims to stay safe from them. Real-world stalking and harassment can result from cyberstalking in certain situations, which may be extremely dangerous.

**2.1.2 Cyberbullying Effects**

Any kind of bullying, including cyberbullying is linked to physical, emotional health difficulties and academic difficulty. Teens subjected to cyberbullying are more likely to suffer from depression, suicidal thoughts or attempts, criminal behaviour, and drug abuse (Yang,et al., 2014).

Cyberbullying can have several negative consequences for both the bully and the victim. Some of these consequences include:

(i). Anxiety and Depression: Anxiety, sadness, and a loss of self-esteem are all potential outcomes of cyberbullying. Victims of cyberbullying may have symptoms of anxiety and despair due to the abuse they have suffered. This can lead to decreased school performance, social withdrawal, and even suicide (L. Yang et al., 2014)

(ii). Suicidal Ideation: In some cases, victims of cyberbullying develop suicidal thoughts. Cyberbullying victims may contemplate or attempt suicide because they feel hopeless and powerless.

(iii). Substance Abuse: Cyberbullying can lead to substance abuse. Victims of cyberbullying may self-medicate with drugs or alcohol to cope with the emotional pain caused by bullying (Alim, 2016). In addition, Bullies may engage in substance abuse to carry out their bullying, such as using drugs or alcohol to lower their inhibitions and make it easier to bully someone.

(iv). Relationship Problems: Relationship issues can be worsened by cyberbullying. Cyberbullying victims may find it challenging to establish relationships because of the emotional trauma they have suffered from bullying.

(v). Academic Problems: Cyberbullying can cause academic problems. Emotional discomfort caused by cyberbullying can make it difficult for the victims to concentrate or participate in school activities.

(vi). Social Isolation: Cyberbullying can lead to isolation in some cases. Victims may withdraw from social activities and isolate themselves to cope with the emotional trauma they have experienced (Cassidy et al., 2013)

(vii) Physical Health Problems: Cyberbullying can lead to physical health problems. Victims of cyberbullying may experience headache, stomachache, and sleep problems due to bullying (Alim, 2016).

Cyberbullying victims at schools who had been harassed, experienced poorer results, including depressive symptoms, thoughts of suicide, and self-injury that needed medical care (Cassidy et al., 2013). Additionally, studies show that cyberbullying has a more significant harmful impact on children than traditional forms of bullying (Cassidy et al, 2012). There are many similarities between cyberbullying and conventional bullying; however, the emotional and psychological show of power in cyberbullying is unlimited and occurs regularly for 60% of people using the internet (Brochado et al., 2016). Despite the lack of a physical attack, it can cause significant damage to the reputation, mental well-being, and social connections of a child or teenager.(Alim, 2016)

## 2.2 Cyberbullying on Social Media

Younger Americans, particularly those in their teens and early twenties, increasingly use social media to abuse each other (Desai et al., 2021). These social media platforms enable teens to send images and posts to connect with friends, and they are also a breeding ground for cyberbullying. It has been observed that the most common places where cyberbullying occurs are on social media platforms like Facebook, Instagram, and Snapchat. Cyberbullies can bring harm to others while maintaining their anonymity on social networking websites, which gives them the ability to do so without fear of being caught. Comments on a child's photo might be public or private, and it is also possible for images and comments to spread quickly through social media. In addition, the availability of smartphones and other mobile devices at home means that children have access to social media round the clock (i.e., iPad, laptops and computers, gaming systems).

Cyberbullying is more likely to happen when teenagers spend more time on social media. Unfortunately, despite the negative consequences, many people continue to use social media because of their social media addiction.

**2.2.1 Twitter and Cyberbullying**

With millions of users all around the world, Twitter has quickly grown to become one of the most popular social networking sites. The platform allows users to share their views on various topics, including politics, brands, goods, and celebrities. Unfortunately, however, the anonymous nature of Twitter has also led to increased cyberbullying activities (Balakrishnan et al., 2020). Studies have shown that Twitter is often used to spread hateful and abusive comments, especially towards women and minority groups. In addition, Twitter has been used to bully and harass individuals, leading to feelings of anxiety and isolation. The rise in cyberbullying activities on Twitter is a cause for concern, and it is essential to take steps to prevent cyberbullying from taking place.

Twitter and other social media platforms have become increasingly popular, contributing to a rise in instances of cyberbullying. Twitter is a popular platform that allows users to post messages known as tweets to their followers. It is used daily by more than 100 million individuals, and tweets are posted an average of 200 million times. The use of Twitter necessitates the utilisation of some specialised features, including content, hashtags, and mentions (Desai et al., 2021).

Since Twitter's goal is to provide a haven for free speech, users are encouraged to report any offensive material they encounter. These policies are being implemented to increase user safety and happiness and lessen the risk of cyberbullying and other forms of harassment.

## 2.3 Text Classification

Text classification is the process of automatically assigning a category to a piece of text. This can be done in several ways, including using machine learning algorithms (Zhao et al., 2016). It can be used for various tasks, including sentiment analysis, spam detection, and customer segmentation. Sentiment analysis determines whether a piece of text is positive, negative, or neutral.(Gaydhani, et al., 2018). This can be used to gauge public opinion on a topic or to understand customer sentiment. Spam detection identifies spammy content, such as unsolicited emails or comments on a blog post, it can be used to protect websites from spam or filter out unwanted content. Customer segmentation is the process of dividing customers into groups based on shared characteristics, it can be used to target marketing messages or to better understand the customer base. Text classification can be used to improve businesses in several ways; It helps understand customer sentiment which can be used to improve customer service or target marketing messages. It can also help identify spam content, which can protect websites from spam and unwanted content (Mirończuk & Protasiewicz, 2018)

**2.3.1 Text Analytics and Sentimental Analysis**

Text analytics is a process of analyzing unstructured text in order to extract meaningful information. Sentiment analysis, a subfield of text analytics, examines the sentiment of the text or the emotions that are expressed (Masood Khan & Rahat Afreen, 2021). The purpose of conducting sentiment analysis is to gain a deeper understanding of the emotions that are conveyed in a piece of writing by identifying patterns and trends in words used. On the other hand, text analytics analyses the meaning of the words and their syntax (Loshin, 2013).

Significant differences which exist between sentiment analysis, and text analytics are as follows:

Text analytics and sentiment analysis are two related but distinct fields of study. Text analytics is a relatively new field that focuses on using computers to extract meaning from large volumes of text. In contrast, sentiment analysis is a more established field that uses statistical methods to analyze the emotions expressed in text (Kabir et al., 2018). Both areas are concerned with extracting information from text, but they differ in their approach and focus. Text analytics is more concerned with understanding the overall meaning of a text, while sentiment analysis is more focused on understanding the emotions expressed in a text (Bisio et al., 2017). This analysis of textual datasets mimics the way in which the human brain processes language. It recognises the parts of speech, determines the relationships among the words and concepts, and then corrects any errors it finds on its own. Regardless of the size of the database or the number of tweets, it is capable of doing both types of analyses simultaneously.

## 2.4 N-Gram

A continuous string of n pieces collected from a particular section of text is referred to as a "n-gram," and it is denoted by letters, syllables, words, or base pairs. In speech recognition, acoustic n-grams are used to model the acoustic properties of spoken utterances. In natural language processing, linguistic n-grams are used to model text's syntactic and semantic properties (Gaydhani et al., 2018). N-grams can be usefully applied in many ways, including statistical language modeling, speech recognition, pattern recognition, data compression, information retrieval, and text mining (Lopez-Gazpio et al., 2019). One way to create n-grams is to use the Natural Language Toolkit (NLTK) library. NLTK is a powerful toolkit for Natural Language Processing in Python. With NLTK, n-grams can be generated from text data.

The way a word is used in phrases with other words around it is critical for a machine to comprehend what it means, and N-gram is one method for doing so. As an example, Brad had to book our tickets today, whereas I had to read this book. A verb, "booking a vacation," refers to planning a trip. In the second instance, the word "book" is being used as a noun, which refers to a book or object. Because of our natural awareness to language cues, we are born with the mental capacity to distinguish between verbs and nouns. On the other hand, robots must study the context of a target word to interpret these cues. With N-grams, context may be found in texts by analysing the words that precede and follows them. 'I care, care less, not welcomed, not unwelcome...' are all examples of bigrams. N-grams are language models that are effective, and are used in natural language processing to capture the context and meaning of words. They are especially beneficial for capturing words applied in negative contexts. Getting a better grasp on the text's true meaning by documenting its larger context and teaching computers to pick up on its linguistic cues (McDonald et al., 2015)

#### **2.5 Context and Sentiment with N-Gram**

The context of an n-gram is the other items in the sequence surrounding it. In most cases, you can tell the tone and sentiment of a text only by looking at the surrounding text, for example; if the N-gram is “I care” then the context might be a caring motive. This would convey the sentiment of the text as positive. Although it creates some less significant texts, N-gram must be used in conjunction with other approaches in order to identify key terms.

N-grams are a powerful tool for machine learning and natural language processing. They can be used to represent the text data in a way that is easy to use for machine learning algorithms. They can also be used to determine the sentiment of the text (Dey et al., 2018).

In sentiment analysis, the N-gram approach is utilised to evaluate the overall tone of a piece of writing or document. For example, an N-gram is distinguished by its n-gram count. Unigrams, bigrams, and trigrams are n-grams with numerous portions of one word.

## 2.6 Text Classification Algorithm

Classification algorithms are important in the field of machine learning. There are so many of them that picking one to utilize can be somewhat complicated with deciding. Besides the various types of classifiers already in existence, new algorithms are being proposed regularly with new applications in mind. This section discusses some notable text classification algorithms or machine learning algorithms.

**2.6.1 Naive Bayes Classifier**

It is a machine learning algorithm that is based on the principle of maximum likelihood, which estimates the probability of an event occurring based on its observed frequency. The key assumption of the algorithm is that all features are independent of each other (F. J. Yang, 2018). This assumption is called the independence assumption or the naivete assumption.

The Naive Bayes technique is a supervised learning algorithm for classifying data based on the Bayes theorem. With an extensive training set, it is used mainly for text classification. With the Naive Bayes classifier, it is easy to create rapid machine learning models that quickly make accurate predictions. Predicting an object's probability is what this classifier is all about. The application of Naive Bayes Algorithm include; spam filtering, sentiment analysis, and article classifications.

**2.6.2 Support Vector Machine**

Classification and regression problems may both be solved using support vector machine (SVM), a supervised learning technique. To optimize the margin between classes, the method uses discriminative classifiers (Cervantes et al., 2020). The algorithm has been widely used in many applications, including facial recognition, text classification, and handwriting recognition, and it can be adapted to different types of data. The SVM algorithm maps data to high-dimensional feature space and then finds the hyperplane that best separates the data points. The hyperplane is defined by a set of support vectors, which are the data points closest to the hyperplane (Cervantes et al., 2020).

Support vector machines can successfully classify data that are multidimensional. Nearest neighbour employs several classes of training data to establish the optimum hyperplane, whereas SVM maximises classification margin instead. Instead of having to do an expensive and time-consuming computation of similarity over several feature space dimensions (Muneer et al., 2020).

**2.6.3 Random Forest**

Random forest is a supervised classification and regression algorithm. This algorithm constructs a collection of decision trees out of the training data by using a portion of it that is chosen at random (Zhang et al., 2021). The final prediction is made by aggregating the predictions of all the individual decision trees. Overfitting and noise issues are less likely to occur when using this approach. However, the training process is long and computationally intensive. (Desai et al., 2021).

**2.6.4 Decision Tree**

A decision tree algorithm is a machine learning approach that divides a dataset into subsets based on specified parameters. It is a method of supervised learning that does not rely on parameters and can be used for classification and regression problems (Charbuty & Abdulazeez, 2021). The main idea behind a decision tree algorithm is to map out all possible solutions to a problem and then choose the best one based on certain conditions. This is similar to how humans make decisions by considering all the possible options and then selecting the best based on specific criteria. The decision tree algorithm is a recursive algorithm, which means it can be used to solve problems that can be divided into smaller sub-problems. The approach is simple to describe and has a limited number of hyperparameters; however, it is ineffective when applied to datasets with several classes or a limited dataset with possible overfitting (Muneer et al., 2020).

**2.6.5 CatBoost Classification Algorithm**

CatBoost is a model that is based on boosting. Boosting consists of iteratively training models on weak learner models (i.e., learners with low accuracy), which are fitted to each other, and then combining the best performers at each step (Muhammed et al., 2020). This means that it will start by fitting a series of weak learners and combining them into one strong learner by determining the one with the best accuracy. This process is repeated until convergence, when no more accuracy improvements are possible. The CatBoost classifier is a relatively new classification method. It has been proposed as an alternative to AdaBoost and XGBoost. One of the main applications of CatBoost is text classification. As with any classification model, the CatBoost classifier can be trained on a corpus of texts and will learn how to classify them correctly. This makes it useful for text classification tasks such as sentiment analysis, where the words in a sentence can reveal whether the sentiment expressed is positive or negative. Another use case is in bioinformatics, where DNA sequence alignment and analysis tasks are required. The CatBoost classifier would be able to classify sequences based on patterns or features. It’s also worth mentioning that besides these two examples, there are many other potential uses of this algorithm.

CatBoost is an efficient classification model. With its ability to classify large datasets in a reasonable amount of time, it is a viable alternative to traditional supervised classification algorithms (Jabeur et al., 2021). Tanha et al. (2020) confirms it is an algorithm created by Yandex, it is open source and free to use, because of its resourcefulness, it reduces the requirement for extensive hyper-parameter adjustments and the risk that results have been overfit, resulting in more generalised models. It is possible to adjust parameters like the learning rate, regularisation, the depth of the trees, the fold size, and the bagging temperature in CatBoost.

Compared to other machine learning techniques, it supports a wider range of descriptive data formats. When applied to multi-class, imbalanced datasets as well as huge datasets, an empirical analysis has shown that the logitboost and CatBoost algorithms are superior to other boosting techniques(Tanha et al., 2020)

## 2.7 Survey of Related Works

In an effort to find a solution to the issue of cyberbullying, previous studies have been utilising text-mining algorithms. In the course of their investigation, text mining was used to identify instances of spam, harassment, and threats.

According to Reynolds et al., (2011) who used supervised machine learning to detect cyberbullying. A question-and-answer site called formspring.me serves as the source of this study's data, allowing users to submit answers and have them voted on by other users. In addition, user profiles were included as part of the generated data set. There were a total of 20 training files and 10 test files, in addition to the 10 training files and 10 test files that were already there. The data were separated into cyberbullying and non-cyberbullying categories with the help of Amazon Mechanical Turk. After the dataset had been labelled, it was determined that there was a class discrepancy within the dataset. 173 of the 1219 messages were found to be cyberbullying. Text-based components were used in the process of developing the C4.5 decision tree model. After compiling a list of cuss words and insults and sorting them from most severe to least severe, the algorithm was able to attain an accuracy of 78.5 percent overall in its predictions.

An approach Dadvar and De Jong, (2012) put out includes factors such as user characteristics and post-harassment behaviour in the development of a model for detecting online bullying. They recommended conducting a cross-system study in their research proposal. This study would analyse the participants' activity throughout their online social networks in order to detect instances of cyberbullying and to prevent its occurrence. In addition to that, their investigation included areas such as vocabulary, second-person and third-person pronouns, and gender. While males were responsible for producing 66 percent of the data, females were only responsible for 34 percent of the 2200 hand classified datasets. In order to identify instances of cyberbullying, Dadvar and De Jong used a support vector machine classifier. Classifier training on the male and female posters was done independently, which is why there was a little improvement in accuracy (Dadvar & De Jong, 2012). However, it is important to point out that these studies did not make use of pre-processing techniques for contextual capabilities of the data used.

A method for representation learning called embeddings has been developed to detect instances of cyberbullying. The Bag-of-Words model, commonly known as enhanced bag of words (eBoW), was another method Zhao et al.(2016b) presented. They sent the data into a linear SVM classifier, and the results were determined by using pre-defined offensive phrases as features and assigning weightings to those features. The final representation uses a variety of features, including bag-of-words features, latent semantic features, and bullying characteristics. According to the findings, the semantic bag of words (sBoW) offers superior performance compared to eBoW. In addition, "eBoW" performed better than "sBoW" in terms of results (Zhao et al., 2016b).

Cyberbullying and abusive personality on social media networks were predicted by Chatzakou et al. (2017). As a data source, Twitter's streaming API was used to build a random forest model. Textual, network, and user-based feature extraction approaches were utilised in the study. All 10 cross-validations were done using the same 10-fold scaling. As a result, 90 percent and 81 percent accuracy were achieved using 3 and 4 class classifications.

To detect hate speech Rasel et al. (2018) used text mining and supervised machine learning models to extract and analyse social media comments. Tweets were collected from Data World and classified as hate speech, offensive speech, and none of the above. To decrease the number of input variables, they employed tokenization, n-gram, and term frequency-inverse document frequency (TF-IDF). At the same time, machine learning models such as Random Forest, Logistic Regression, and Support Vector Machines were used to predict hostile remarks. However, when it came to recognising hate speech, offensive speech, and neither category, the SVM model had the weakest performance, whereas other models performed better (Rasel et al., 2018). Using accuracy as the only evaluation metric the random forest classifier, logistic regression and support vector machine attained an overall accuracy of 93%, 82%, and 76% respectively. The SVM model produced lots of misclassification and performed poorly when compared to other classifiers algorithm.

.

As part of the study made by Van Hee et al.,(2018a), researchers looked for cases of cyberbullying by conducting several binary classification tests. Using a series of binary classification tests, the study examined the possibility of automatically detecting cyberbullying by collecting and annotating a cyberbullying corpus in English and Dutch. They made used linear support vector machines with a rich feature set, and they analyse which information sources are most helpful for the detection of cyberbullying. For the identification of cyberbullying, a hold-out test set yielded encouraging results. After optimizing the hyperparameters, F1 scores of 64% and 61% are achieved for English and Dutch.

Machine learning approaches were explored by Zhang et al. (2019) and deep learning models, including Random forest, Decision tree, and deep neural networks. This study analysed 1,395 cyberbullying tweets and 1,395 non-cyberbullying tweets. Using character n-grams, the character 4-gram has been found to be the most accurate predictor. 70% of tweets are utilised as training data for all machine learning models, while the remaining 30% are used as test data. Word2Vec, word n-gram, and character n-gram have documented qualities that can significantly benefit classification. The Logistic regression or Gradient boosting regression tree features and models achieved over 90% for all four assessment criteria.

Using neural networks and support vector machine Hani et al.(2019) suggested a method to identify cyberbullying. A variety of n-gram language models' classifications were examined. As a result, neural Networks achieved 92.8 percent accuracy, while SVMs achieved 90.3 percent accuracy by combining term frequency-inverse document frequency (TF-IDF) and sentiment analysis into their algorithms. Furthermore, it was discovered that the SVM classifier, which obtains an average f-score of 89.8%, is outperformed by the Neural Network classifier, which likewise achieves an average f-score of 91.9% The approach was carried out using a small amount of data.

A supervised approach was built to identify cyberbullying by integrating embedding, sentiment, and linguistic characteristics with Pointwise Mutual Information (PMI) semantic orientation, which Talpur and O'Sullivan (2020) suggested as a framework for generating features from tweets. With the use of Naive Bayes, KNN, Decision Tree, Random Forest, and Suppor vector machine, they solved a multi-class classification problem with unbalanced data using the extracted features. Because the data in this article has an unbalanced class, the authors relied on Area Under Curve (AUC) as their primary performance assessment parameter for binary classification. Experiments with poor results were taken off the list so that the best results could be compared across each layer of features that add value to the classifier's performance. The validity of the models were tested using a 10-fold cross-validation approach. They classified tweets into one of four severity levels: none, low, medium, or high, based on the severity of the messages. The overall performance of the base classifier was marginally enhanced by using the SMOTE parameter, which dealt with class imbalance distribution. With an area under the curve (AUC) of 0.971 and an F-measure of 0.929. Random Forest has been shown to be the most effective classifier in binary classification settings. This finding is comparable to that of multi-class classification (Talpur & O’Sullivan, 2020)

Rezvani et al.(2020), using photos, metadata, and textual content created around an image, were able to identify cases of cyberbullying. A Deep Neural Network focusing on text content and a second Neural Network to identify contextual information were used to uncover and improve cyberbullying detection. By combining the findings of the two networks, they were able to use textual and contextual data from some sources. Their proposed detection method performed optimally with contextual data. Using the neural network model with textual and contextual data, the model had an accuracy of 85% , other evaluation metric such as precision, and f1 score records 87% and 85% respectively.

Muneer and Fati (2020) compiled a global dataset of 37,373 unique tweets from Twitter to investigate this problem. Their objective was to collect tweets, classify them using text analysis techniques based on specified keywords, and classify them as either offensive or non-offensive. This would allow them to detect instances of cyberbullying. Support Vector Machine, Logistic Regression, Light Gradient Boosting Machine, Stochastic Gradient Descent, Random Forest, AdaBoost, and Naive Bayes were used to test each feature individually and evaluate their accuracy and were also tested for their ability to classify a global dataset using metrics such as accuracy, precision, recall, and F1 score. A median accuracy of 90.57 percent was reached by LR in the experiments, demonstrating its superiority. Logistic regression had the highest F1 score (0.928), SGD had the highest precision (0.968), and SVM had the highest recall among the classifiers tested 1.0

Mahbub et al.(2021) in this study specifically focused on sexual harassment as a kind of cyberbullying, and the vocabulary of approach words they developed as a result is unique. A Java XML SAX (Simple API for XML) parser was used to extract the textual contents of each chat from the Perverted-Justice website, which provides countless transcripts of chat-based talks between decoys posing as teenagers and convicted sexual offenders in the United States of America. The Stanford CoreNLP package for Java analyzed the text and extracted nouns, noun phrases, verbs, and verb phrases. The datasets were analysed using the J48 decision tree classifier, JRip rule-based classifier, and Naive Bayes classifier, all of which are accessible in the WEKA suite. Cyberbullying content was indicated by a "Yes" or "No" label, respectively. Improvements in all three classifiers were the same. With an improvement in J48 precision from 80.07 percent to 81.60 percent, the model's recall for the class label "Yes" remained constant at 0.54 in both tests. The class label "Yes" improved accuracy from 0.80 to 0.86, with JRip and Naive Bayes methods measuring similarly to J48.

## 2.8 Research Gap

According to previous studies, the following gaps in existing studies on cyberbullying detection can be observed: First, in the early stages of cyberbullying detection, social media texts are heavily exploited. New text recognition systems are necessary to capture additional semantic meaning and context in text. Another issue is that only a few studies can detect cyberbullying when they employ unbalanced datasets from various social media platforms. Despite research on the issue of cyberbullying, the CatBoost classification algorithm is yet to be evaluated in conjunction with the N-gram language model to determine the sentiment of texts received over social networks.

**CHAPTER THREE**

## METHODOLOGY

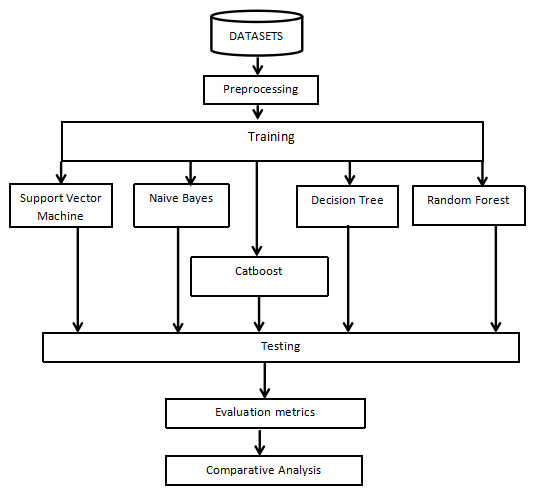
## 3.1 INTRODUCTION

In this section, the process for detecting cyberbullying is outlined in detail, along with the datasets used. In particular feature extraction, feature selection approaches, and performance assessment metrics employed are described. In addition, in order to increase the accuracy of detecting occurrences of cyberbullying in texts, this section makes use of the CatBoost classifier in conjunction with N-gram. Using text classification methods such as the CatBoost classifier and other classifier algorithms, this research pulls information on cyberbullying words and non-cyberbullying words from the linguistic factors of the data. The goal is to identify the features that makes these two types of words different. The performance of the CatBoost classifier will be compared to the performance of the other classifier algorithms that have been discussed and applied in the detection of cyberbullying, and evaluation metrics will be used to conduct this comparative study of the classification algorithms.

## 3.3 Research Framework

The methodological framework depicted in Figure 3.1 is employed in this dissertation. The framework is made up of several stages, and the results of each stage are used as inputs for the next stage. Collecting datasets was the first step, followed by preprocessing, training using algorithms for classification, and testing and prediction. The models accuracy, precision, recall, and F1 score are used to evaluate each classifier.

The Python natural language toolkit and Scikit-learn platforms which has all the resources and capabilities for developing, training and testing the models, were used for implementation.



### Figure 3.1: Research framework

Five machine learning algorithms are selected for this study. They include the CatBoost Classifier, which is the main focus of this study, and four other algorithms, SVM, Naive Bayes, Decision Tree, and Random Forest, which are commonly used in the literature. In addition, they are selected for comparative purposes.

## 3.4 Data Collection

The datasets were obtained from the Mendeley repository (Elsafoury, 2020) and Kaggle website. In order to test whether or not model performance is data-dependent, separate datasets were collected from these data sources.

This research implementation presents a binary classification task due to the fact that there are two different categorization groups, cyberbullying or non-cyberbullying text.

There are a variety of tweets in the dataset that have been preprocessed so that the important textual characteristics may be extracted to get the best model performance from the classifiers.

### Table 3.1: Details of data collected

|  |  |  |  |
| --- | --- | --- | --- |
| **DATA SOURCE** | **TOTAL NO** | **CYBERBULLYING**  **CLASS** | **NON CYBERBULLYING CLASS** |
| Mendeley | 11877 | 5055 | 6822 |
| Kaggle | 115625 | 14766 | 100859 |

**3.5 Balancing the Data**

The issue of imbalance classes in data is a common problem or occurrence in text classification task. The minority class of data often faced by this limitation are mostly the focus area or the subject of the model. This problem arises due to limited data available.

In order to ensure the effectiveness of a model and also reduce over-fitting and under-fitting issues caused by this imbalance classes of data, it is very important to solve the problem of imbalance before creating a machine learning model.

There different method of solving imbalance classes in machine learning which are:

* Oversampling
* Under-sampling
* Synthetic minority oversampling techniques(SMOTE) among others.

Oversampling is commonly used to increase the number of the minority class in order to create a balance with the majority class, the oversampling techniques engages data duplication algorithm which is not good enough in cases of highly imbalanced data.

The under-sampling technique reduces the number of the majority class in order to achieve a balance in both classes thereby leading to loss of data. The oversampling and under-sampling techniques are limited in the case of fairly and highly imbalanced classification. In order to solve this limitation. The SMOTE techniques helps by oversampling the minority class through the use of synthetically generated datas, in this methods, no duplicates are created and there are no data loss.

For this study, the SMOTE technique was used to solve the problem of imbalance classification.

## 3.6 Text Preprocessing

Text preprocessing turns raw text into a form that machine learning algorithms can use. The steps in text preprocessing usually involve cleaning up the text, tokenizing it, and converting it into a numerical representation. Cleaning the text usually involves removing punctuation, whitespace, and other non-letter characters (Kadhim, 2018). Tokenizing the text breaks it up into individual words or phrases. And converting the text into a numerical representation can be done using a bag of words approach or word embeddings. The python natural language tool kit (NLTK) was instrumental in the general text preprocessing, tools such as word\_tokenize, word\_count, and stopwords from nltk.corpus were extracted from the nltk library.

As a natural extension of the bag of words technique and the preceding steps, this study uses bags of n-grams to extract features. N-grams are groups of words that occur together in a text. N-grams can analyze the patterns of word use in a text and identify keywords and phrases that are important to a text's meaning (McDonald et al., 2015). The machine learning model can now use different preprocessed datasets for training and testing with any algorithm of choice, the values of n-grams used in this study are 1, 2, and 3 for one-gram, bi-gram and tri-gram respectively. Frequently used words in the datasets were analysed based on the n-gram words used.

## 3.7 Evaluation Metrics

A variety of evaluation metrics was used to examine whether or not the models are effective at distinguishing between messages that represent cyberbullying and texts that do not. To completely comprehend the performance of various models, standard assessment methodologies must be applied. The following are some metrics that can be used with machine learning classifiers to evaluate social media networks (such as Twitter) in relation to cyberbullying:

**(i) Accuracy**

Accuracy measures how well a machine learning model predicts the correct labels for a given set of data. It is one of the most important metrics to consider when evaluating a model, as it gives us a clear indication of how well the model is performing. Several factors can affect accuracies, such as the type of data used, the size of the training dataset, and the complexity of the model. However, in general, we can say that a model is accurate if it makes correct predictions more often than it makes incorrect predictions.

Regarding cyberbullying detection in this study, accuracy is measured by the number of accurate predictions made for true cyberbullying and true non-cyberbullying occurrences. In this instance, accuracy is calculated by dividing the number of correctly predicted samples by the total sample count. In terms of cyberbullying and non-cyberbullying predictions, true positives and true negatives represent the number of correct predictions, respectively, while false positives and false negatives represent the number of incorrect predictions.

**(ii) Precision**

Precision is a performance measure in machine learning that looks at how well a model predicts. Precision measures how many of the predicted values are correct. It is a ratio of the number of true positives to the total number of predicted positives. It is often used in conjunction with recall, which correctly represents the percentage of positive values predicted by the model. Combining these two metrics can give you a clear grasp of how accurate a model is.

The results of how many instances of cyberbullying words evaluated to be positive by each classifier can be calculated using precision. This considers both the number of correctly predicted cases of cyberbullying that are true positives and the number of incorrectly predicted cases of cyberbullying that are false positives. The cyberbullying detection model receives a lower score for its precision whenever a more significant number of negatives are incorrectly labelled as positives. It is calculated by dividing the number of true positive results that the model actually achieved by the total number of true positive and false positive results that were predicted by the classifier.

**(iii) Recall**

A model with high recall will have a low false-negative rate, meaning that it can correctly identify positive examples. Conversely, a model with low recall will have a high false-negative rate, meaning that it is often unable to identify positive examples correctly. The recall is often used alongside other measures such as precision and accuracy to get a complete picture of the performance of a machine learning model. In some cases, recall may be more important than precision, especially in applications where false negatives are more harmful than false positives (e.g., cancer detection). Precision may be more important than recall (e.g., spam filtering) in other cases.

In this study, the evaluation of all instances of positive cyberbullying text, both true positives and false negatives, is accomplished. Therefore, the recall score will decrease according to the number of positives the classifiers miss throughout the process of detecting cyberbullying texts, leading to an increased number of false negatives.

**(iv) F1 Score**

The F1 score is often used to compare different models. A model with a higher F1 score is usually considered to be better than a model with a lower F1 score.

F1 score can be calculated using the following formula:

The F1 score ranges from 0 to 1, with a higher score indicating a better model. While developing cyberbullying detection texts, it was necessary to examine the Harmonic Mean of each classifier to provide a single score that accounted for both precision and recall. It is based on the F1 score that determines which model is the best. A classifier's accuracy (the number of times it classifies correctly) and robustness are measured using this metric.

**CHAPTER FOUR**

# 4.0 RESULTS AND DISCUSSION

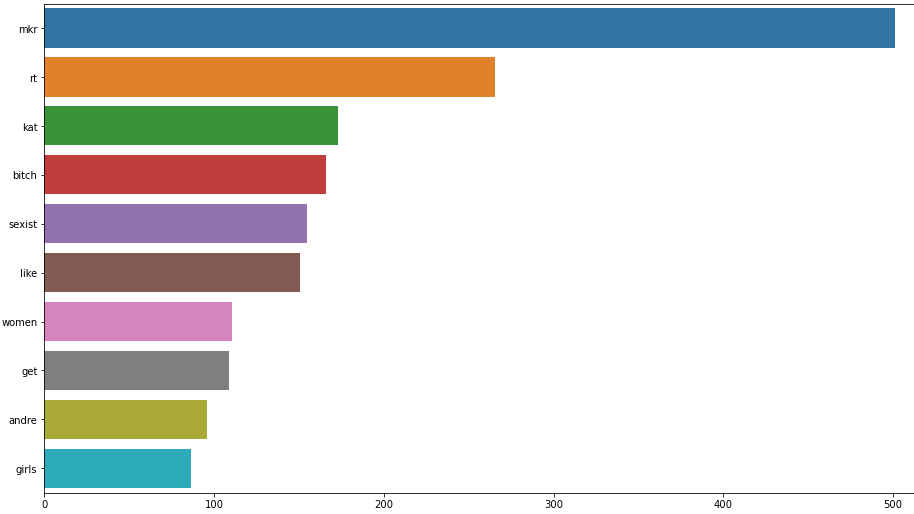
## 

## 4.1 Introduction

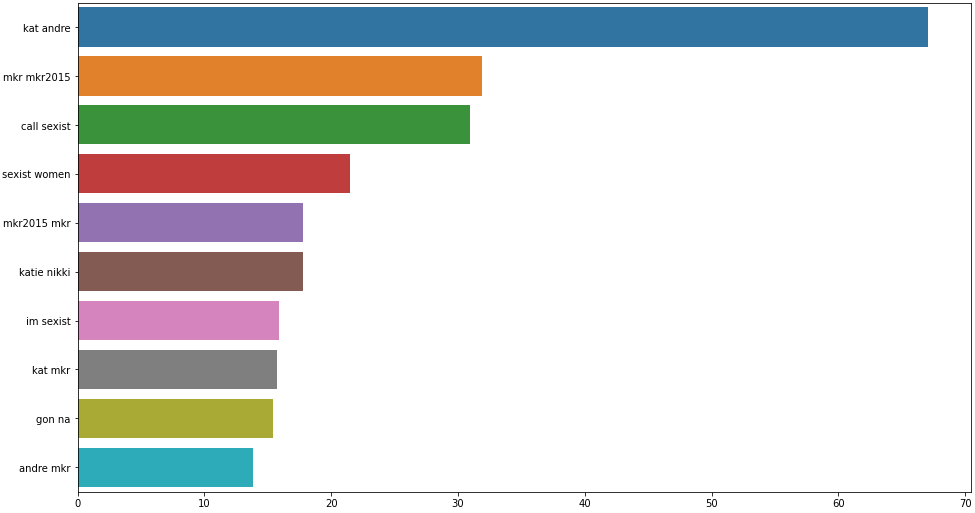
Using accuracy, precision, and recall f1 scores, the support vector machine, decision tree, random forest, and naive bayes model were compared with the CatBoost classifier; these metrics measure how well each of the algorithms were able to detect texts with instances of cyberbullying that had already been trained.

**4.2 Results from N-Gram Analysis**

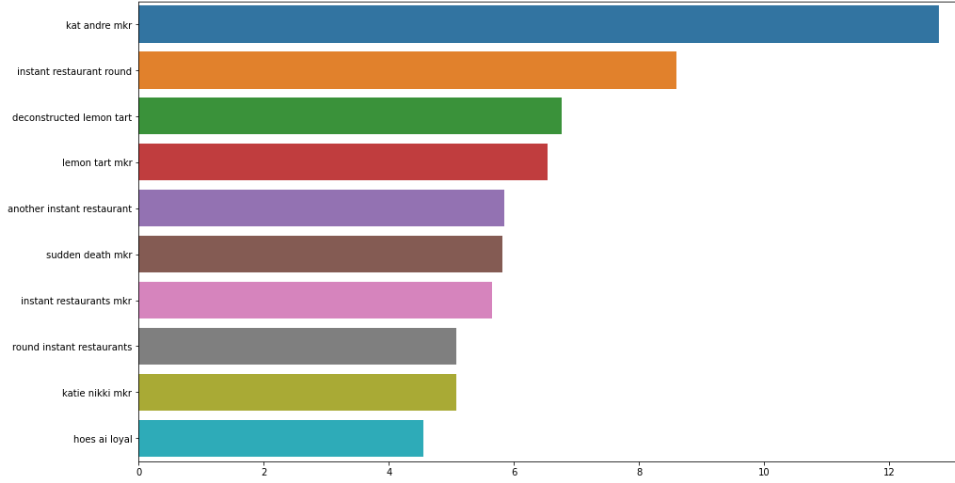
A language model incorporating n-gram was created by counting the number of times each unique n-grams (where n=1,2,3) appears in the textual datasets. An n-gram model predicts the most probable word that might follow a word sequence. Its a probabilistic language model that is trained on a corpus of text. Figures 4.1, 4.2 and 4.3 depict the frequently used combination of unigram, bigram and trigram respectively in the dataset text corpus.



### Figure 4.1: unigram visualization for text corpus



### Figure 4.2: bigram visualization for text corpus



### Figure 4.3: trigram visualization for text corpus

## 

## 4.3 Datasets Distribution

The distribution of the dataset collected from mendeley is as follows:

There were a total of 11877 tweets with labels.

Tweets with no bullying: 6822

Number of bullying-related tweets: 5055



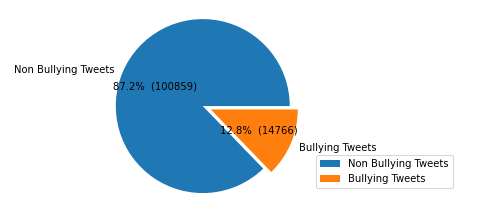
### Figure 4.4: Mendeley dataset distribution

The distribution of the dataset collected from kaggle is as follows:

There were a total of 115625 tweets with labels.

Tweets with no bullying: 100859

Number of bullying-related tweets: 14766



### Figure 4.5: Kaggle dataset distributions

## 4.4 Results

Exploratory data analysis, which involves data cleansing and pre-processing, was used to examine the data. 70% of the datasets were used for training, while 30% was used for testing.

The results of the datasets show that they are imbalance, the bullying tweets are in the minority class. Thus, the interpretation of the performance measures will be with respect to the minority class the bully tweets which is the focus of the study.

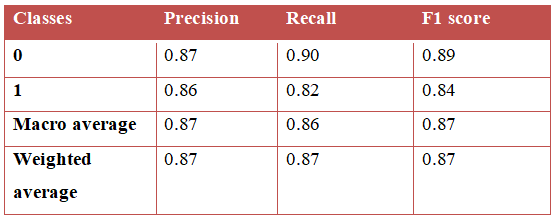
**4.4.1 Training and Testing Using Mendeley Dataset**

Tables 4.1 - 4.5 presents the performance results of SVM, Decision tree, Random forest, Naive bayes and CatBoost respectively on the Mendeley dataset. Thetable show the precision, recall, and f1-score measures of the models. CatBoost and SVM having marginally better results that the other models according to the Table of results.

The classes 0 and 1 in tables 4.1 - 4.5 represents Non-Cyberbullying texts and Cyberbullying text respectively.

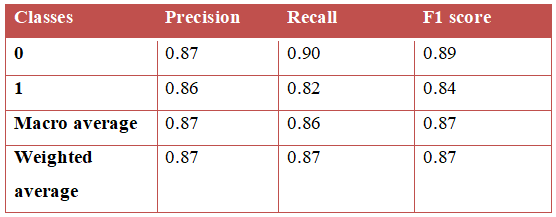
Accuracy = 87%

### Table 4.1 Results of SVM model



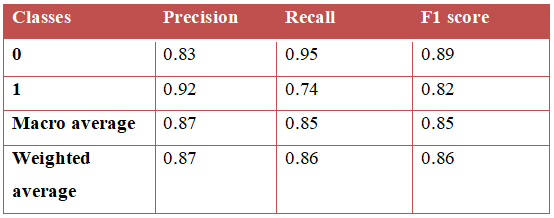
Accuracy = 86%

### Table 4.2 Results of Decision Tree model



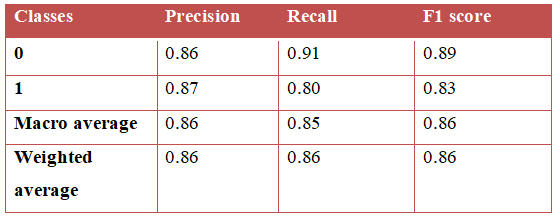
Accuracy = 86%

### Table 4.3 Results of Random Forest model



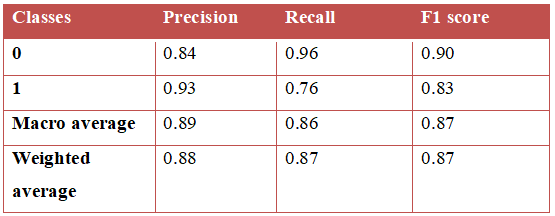
Accuracy = 86%

### Table 4.4 Results of Naive bayes model



Accuracy = 87%

### Table 4.5 Results of CatBoost classifier model



**4.4.2 Observation**

The Mendeley dataset has a pretty uniform distribution of data prior to preprocessing. CatBoost and SVM outperformed the other classification algorithms in this study, with an average f1 score of 0.87 and an accuracy rate of 87%.

**4.4.3 Training and Testing Results Using Kaggle** **Datasets**

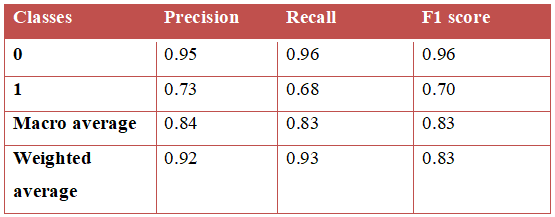
Tables 4.6 - 4.10 presents the performance results of SVM, Decision tree, Random forest, Naive Bayes, and CatBoost, respectively, using the Kaggle dataset. The accuracy of the models shows that CatBoost classifier has marginally better accuracy than other models.

The classes 0 and 1 in tables 4.6 - 4.10 represents Non-Cyberbullying texts and Cyberbullying text respectively.

### 

### Table 4.6 Results of SVM model

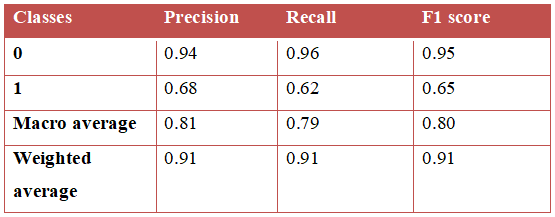
Accuracy = 93%



### 

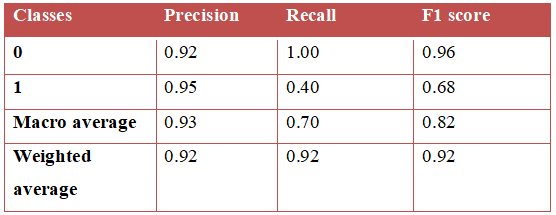
**Table 4.7 Results of Decision Tree model (Kaggle dataset)**

Accuracy = 91%



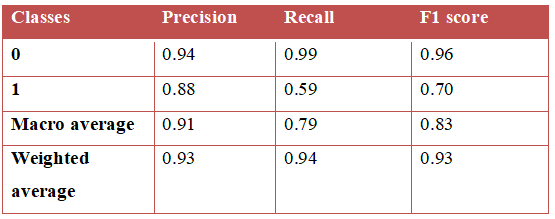
### Table 4.8 Results of Random Forest model

Accuracy = 93%



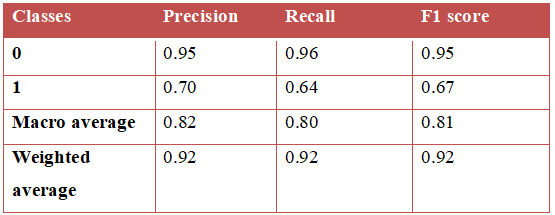
### Table 4.9 Results of CatBoost classifier

Accuracy = 94%



### Table 4.10 Results of Naive Bayes

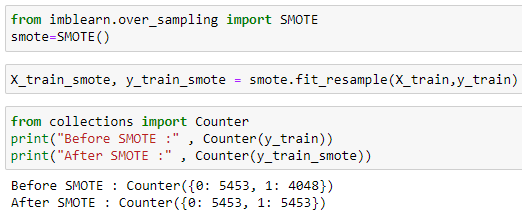
Accuracy = 92%



**4.4.4 Observation**

The dataset that was used from kaggle is far bigger and more unevenly distributed than the datasets that were acquired from mendeley. After training and testing, the CatBoost classifier achieved a 94% accuracy and 0.93 weighted average f1 score, placing it above all other classifier considered in this work.

**4.5 Results and Observation of Using Balanced Mendeley Dataset**

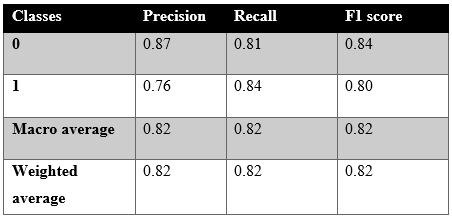
The cyberbullying and non-cyberbullying texts were equal at 5,453 in each class after balancing the dataset with SMOTE approach.

Tables 4.11 – 4.15 shows the results of SVM, Decision tree, Random forest, Naive Bayes, and CatBoost on the balanced Mendeley dataset. The accuracy, recall, and f1-score measurements of the models are shown in the tables. The accuracy of the models are 82%, 76%, 79%, 86%, and 86%, respectively.

CatBoost has the same accuracy score as Naïve bayes, a weighted average f1 score of 0.86 and a precision of 86%. They outperformed the other classification algorithms employed in this study, demonstrating that model performance is data dependant as a result of the modifications made to the datasets.

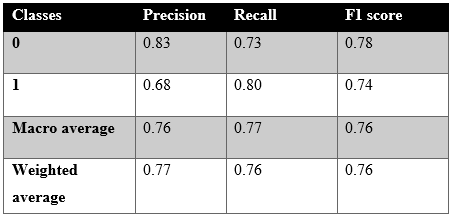
### Table 4.11 Results of SVM model

Accuracy = 82%



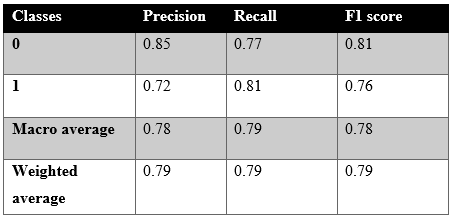
### Table 4.12 Results of Decision Tree model

Accuracy = 76%



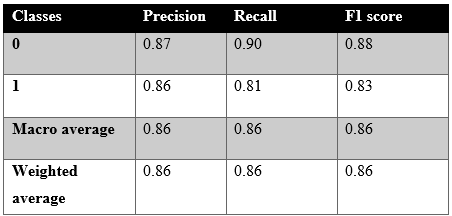
### Table 4.13 Results of Random Forest model

Accuracy = 79%



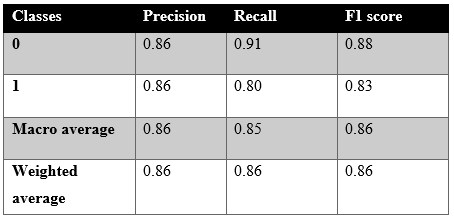
### Table 4.14 Results of Naive Bayes model

Accuracy = 86%

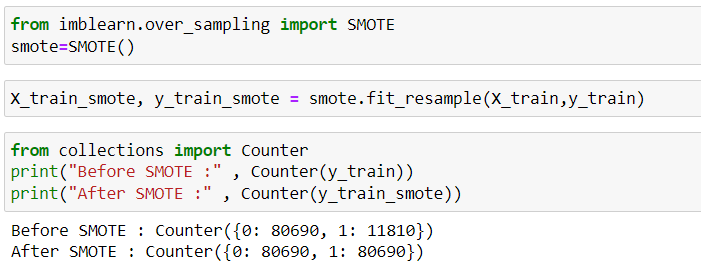


### Table 4.15 Results of CatBoost classifier model

Accuracy = 86%



**4.6 Results and Observation Using Balanced Kaggle Dataset**

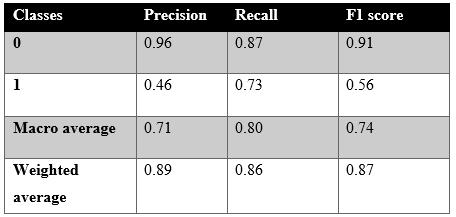
SMOTE was used to balance the dataset, and the cyberbullying and non-cyberbullying classes were equal on 80,690 texts in each class.

SVM, Decision tree, Random forest, Naive bayes, and CatBoost performance results on the balanced Kaggle dataset are presented in Tables 4.16 - 4.20. The accuracy, recall, and f1-score measurements of the models are shown in the tables. The models' accuracy are 86%, 81%, 86 %, 91, and 87%, respectively.

The weighted average f1 score for CatBoost is 0.88, and its accuracy is 87%, with SVM doing better on accuracy but CatBoost performing better on other parameters. This proves that the models are data-dependent. CatBoost beat all other classifiers before and after balancing the dataset by dealing with the major subject of this study (Cyberbullying texts), which is the minority class, and performing better in accuracy and other metrics.

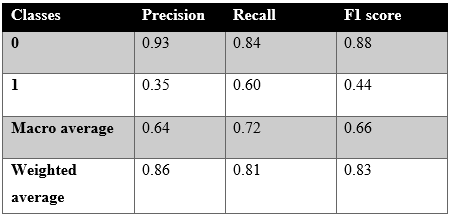
Accuracy = 86%

### Table 4.16 Results of SVM model



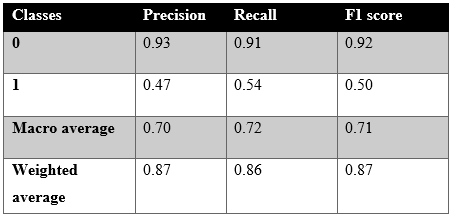
Accuracy = 81%

### Table 4.17 Results of Decision Tree model



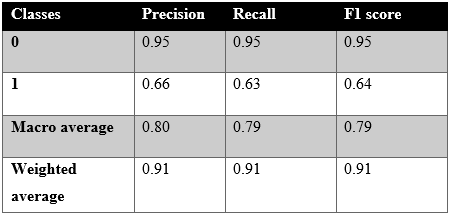
Accuracy = 86%

### Table 4.18 Results of Random Forest model



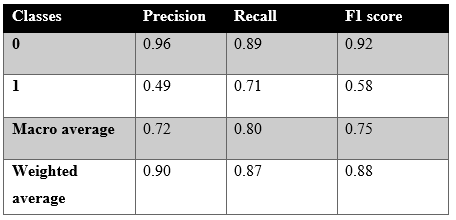
Accuracy = 91%

### Table 4.19 Results of Naive bayes model

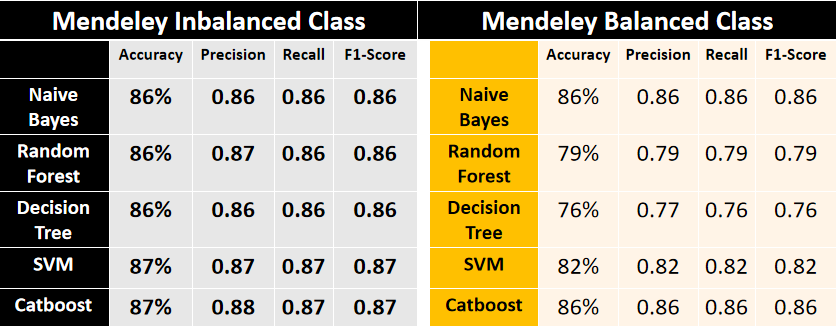


Accuracy = 87%

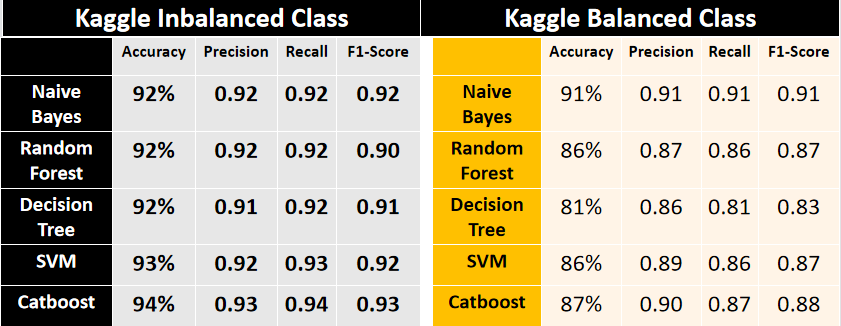
### Table 4.20 Results of CatBoost classifier model



### Table 4.21 Mendeley Imbalanced and Balanced result

****

### Table 4.23 Kaggle Imbalanced and Balanced result

****

**4.7 Conclusion**

The results of the models shows that CatBoost classifier, which is the main focus of this study is equally a good classification model for cyberbullying detection. Although marginally, the CatBoost Classifier outperformed the other models considered in most cases. It also performed marginally with SVM and Naive Bayes in a few instance on the balanced and imbalanced dataset.

Being an efficient algorithm in terms of accuracy, which is a reasonable metric to measure the detection of cyberbullying, CatBoost classifier for cyberbullying detection will be most suitable for online realtime classification of social media text.

**CHAPTER FIVE**

# 5.0 SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

## 5.1 Research Summary

As the Internet has grown, cyberbullying has become a widespread issue that has serious effects on individuals and communities. The goal of this study is to create an automated system for identifying instances of cyberbullying that may be used to reduce its harmful consequences.  
In order to develop a model capable of identifying cyberbullying, textual data pertaining to cyberbullying was trained and tested using the ngram language and the CatBoost classifier. The datasets used was obtained from kaggle and mendeley data repository. The datasets collected was first preprocessed in order to ensure that the data is cleaned and there are no missing values, the n gram language model was used to help the machine learning model used to help learn the context and sentiment of words using uni-gram, bi-gram and tri-gram.

CatBoost algorithm's performance was compared to that of naïve bayes classifier, support vector machine, random forest, and decision tree in order to justify performance. When compared with other classification algorithms taken into account in this study, the CatBoost method performed better across all datasets.

**5.2 Conclusion**

Cyberbullying and non-bullying texts were classified using machine learning algorithms in this study.

Although previous research has engaged the use of some conventional machine learning models such as Support Vector Machine, Naive Bayes, Decision Tree, Logistic Regression, and Random Forest, previous studies also further justify that despite the fact that these conventional classifier algorithms performed well to a degree, there exist certain limitations which include the inability of this conventional model to solve the issues of imbalance classes, high time consumption in processing and training textual data, and others discussed in the review section. As a result, this research used the CatBoost classifier with N-gram to resolve the shortcomings of existing research. The CatBoost classifier algorithm was used owing to its ability to deal with imbalanced classes and categorical variables presented in a binary classification problem. Performance analysis was also carried out on the models used. The performance results of the CatBoost classification algorithm was compared with those of Naive Bayes, Support Vector Machine, Decision tree, and Random forest. The models were evaluated using accuracy, precision, recall, and F1 score metrics.

When applied to the Kaggle dataset, the results of the CatBoost model's classification produced an accuracy of 94%, while using it to the Mendeley dataset resulted in an accuracy of 87%. This accuracy also topped the charts when compared to the other four classification algorithms. Performance of the model varied greatly between the two datasets used for training and evaluation, demonstrating that the success of machine learning models is strongly dependent on the data they are given to work with.

## 5.3 Recommendation

The results of this study are encouraging, and the developed model can be applied to the detect of a real life instances of cyberbullying . Future work on this study's technique will benefit from a larger dataset. As more data becomes available, it will be much easier to spot cyberbullying because cyberbullies are always upgrading their methods of operation. Using more datasets in tandem with the CatBoost will allow for the creation of a more effective system to safeguard users from cyberbullying. This approach may be utilized in both public and private contexts, making it ideal for social networking.

The findings of this research also show that the CatBoost algorithm is superior than more traditional methods for automatically identifying cyberbullying in text. Given how well the CatBoost classification algorithm performed in this study's classification task, it is proposed that this method receives further attention in coming studies. In addition, in order to collect information from all social media platforms where cyberbullying occurs, future research should concentrate more on developing algorithms that can automatically find cyberbullying texts and other cyberbullying instances in audio and video footage published on social media platforms and the internet at large.

# REFERENCES

Alim, S. (2016). Cyberbullying in the world of teenagers and social media: A literature review. In International Journal of Cyber Behavior, Psychology and Learning (Vol. 6, Issue 2, pp. 68–95). IGI Global. https://doi.org/10.4018/IJCBPL.2016040105

Al-Rahmi, W. M., Yahaya, N., Alturki, U., Alrobai, A., Aldraiweesh, A. A., Omar Alsayed, A., & Kamin, Y. bin. (2020). Social media–based collaborative learning: the effect on learning success with the moderating role of cyberstalking and cyberbullying. Interactive Learning Environments. https://doi.org/10.1080/10494820.2020.1728342

Arif, M. (2021). A Systematic Review of Machine Learning Algorithms in Cyberbullying Detection: Future Directions and Challenges. Journal of Information Security and Cybercrimes Research, 4(1), 01–26. https://doi.org/10.26735/gbtv9013

Balakrishnan, V., Khan, S., & Arabnia, H. R. (2020). Improving cyberbullying detection using Twitter users’ psychological features and machine learning. Computers and Security, 90. https://doi.org/10.1016/j.cose.2019.101710

Bisio, F., Oneto, L., & Cambria, E. (2017). Sentic Computing for Social Network Analysis. Sentiment Analysis in Social Networks, 71–90. https://doi.org/10.1016/B978-0-12-804412-4.00005-X

Brochado, S., Soares, S., & Fraga, S. (2016). A Scoping Review on Studies of Cyberbullying Prevalence Among Adolescents. Trauma, Violence, & Abuse. https://doi.org/10.1177/1524838016641668

Cassidy, W., Faucher, C., & Jackson, M. (2013). Cyberbullying among youth: A comprehensive review of current international research and its implications and application to policy and practice. School Psychology International. https://doi.org/10.1177/0143034313479697

Chatzakou, D., Kourtellis, N., Blackburn, J., Cristofaro, E. De, Stringhini, G., & Vakali, A. (2017). Detecting Aggressors and Bullies on Twitter. 26th International World Wide Web Conference 2017, WWW 2017 Companion (2017) 767-768, 767–768. https://doi.org/10.1145/3041021.3054211

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing, 408, 189–215. https://doi.org/10.1016/J.NEUCOM.2019.10.118

Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends, 2(01), 20–28. https://doi.org/10.38094/jastt20165

Cross, D., Barnes, A., Papageorgiou, A., Hadwen, K., Hearn, L., & Lester, L. (2015). A social–ecological framework for understanding and reducing cyberbullying behaviours. Aggression and Violent Behavior, 23, 109-117. https://doi.org/10.1016/j.avb.2015.05.016

Denny, M. J., & Spirling, A. (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. Political Analysis, 26(2), 168–189. https://doi.org/10.1017/PAN.2017.44

Dey, A., Jenamani, M., & Thakkar, J. J. (2018). Senti-N-Gram: An n-gram lexicon for sentiment analysis. Expert Systems with Applications, 103, 92–105. https://doi.org/10.1016/J.ESWA.2018.03.004

Doxing: What it is and how to protect yourself | NortonLifeLock. (n.d.). Retrieved May 28, 2022, from https://us.norton.com/internetsecurity-privacy-what-is-doxing.html

Desai, A., Kalaskar, S., Kumbhar, O., & Dhumal, R. (2021). Cyber Bullying Detection on Social Media using Machine Learning. ITM Web of Conferences, 40, 03038. https://doi.org/10.1051/itmconf/20214003038

Explanation | CyberSecurity Training | www.cybertraining365.com. (n.d.). Retrieved May 28, 2022, from https://www.cybertraining365.com/cybertraining/Topics/Masquerade

Elsafoury, F. (2020). Cyberbullying datasets. 1. https://doi.org/10.17632/JF4PZYVNPJ.1

El Asam, A., & Samara, M. (2016). Cyberbullying and the law: A review of psychological and legal challenges. Computers in Human Behavior, 65, 127-141. https://doi.org/10.1016/j.chb.2016.08.012

Gaydhani, A., Doma, V., Kendre, S., & Bhagwat, L. (2018). Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach. https://doi.org/10.48550/arxiv.1809.08651

Ghori, K. M., Abbasi, R. A., Awais, M., Imran, M., Ullah, A., & Szathmary, L. (2020). Performance Analysis of Different Types of Machine Learning Classifiers for Non-Technical Loss Detection. IEEE Access, 8, 16033–16048. https://doi.org/10.1109/ACCESS.2019.2962510

Hani, J., Nashaat, M., Ahmed, M., Emad, Z., Amer, E., & Mohammed, A. (2019). Social media cyberbullying detection using machine learning. International Journal of Advanced Computer Science and Applications, 10(5), 703–707. https://doi.org/10.14569/ijacsa.2019.0100587

Jabeur, S. ben, Gharib, C., Mefteh-Wali, S., & Arfi, W. ben. (2021). CatBoost model and artificial intelligence techniques for corporate failure prediction. Technological Forecasting and Social Change, 166, 120658. https://doi.org/10.1016/J.TECHFORE.2021.120658

Kabir, A. I., Karim, R., Newaz, S., & Hossain, M. I. (2018). The Power of Social Media Analytics: Text Analytics Based on Sentiment Analysis and Word Clouds on R. Informatica Economică, 22(1). https://doi.org/10.12948/issn14531305/22.1.2018.03

Kadhim, A. (2018). An Evaluation of Preprocessing Techniques for Text Classification Pattern Recognition View project Improvement text classification using log(TF-IDF) with K-NN Algorithm View project. Article in International Journal of Computer Science and Information Security. https://sites.google.com/site/ijcsis/

Lopez-Gazpio, I., Maritxalar, M., Lapata, M., & Agirre, E. (2019). Word n-gram attention models for sentence similarity and inference. Expert Systems with Applications, 132, 1–11. https://doi.org/10.1016/J.ESWA.2019.04.054

Loshin, D. (2013). Data Governance for Big Data Analytics: Considerations for Data Policies and Processes. Big Data Analytics, 39–48. https://doi.org/10.1016/B978-0-12-417319-4.00005-3

Mahbub, S., Pardede, E., & Kayes, A. S. M. (2021). Detection of Harassment Type of Cyberbullying: A Dictionary of Approach Words and Its Impact. Security and Communication Networks, 2021. https://doi.org/10.1155/2021/5594175

Mahesh, K., Gothane, S., Toshniwal, A., Nagarale, V., & Gopu, H. (2021). Cyber Bullying Detection on Social Media using Machine Learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. https://doi.org/10.32628/cseit217381

Muneer, A., & Fati, S. M. (2020). A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter. Future Internet, 12(11), 187. MDPI AG. Retrieved from http://dx.doi.org/10.3390/fi12110187

Mahbub, S., Pardede, E., & Kayes, A. S. M. (2021). Detection of Harassment Type of Cyberbullying: A Dictionary of Approach Words and Its Impact. Security and Communication Networks, 2021. https://doi.org/10.1155/2021/5594175

Masood Khan, A., & Rahat Afreen, K. (2021). An approach to text analytics and text mining in multilingual natural language processing. Materials Today: Proceedings. https://doi.org/10.1016/J.MATPR.2020.10.861

McDonald, G., Macdonald, C., & Ounis, I. (2015). Using part-of-speech n-grams for sensitive-text classification. ICTIR 2015 - Proceedings of the 2015 ACM SIGIR International Conference on the Theory of Information Retrieval, 381–384. https://doi.org/10.1145/2808194.2809496

Mirończuk, M. M., & Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. Expert Systems with Applications, 106, 36–54. https://doi.org/10.1016/J.ESWA.2018.03.058

Muhammed, M. M., Ibrahim, A. A., Ridwan, R. L., Abdulaziz, R. O., & Saheed, G. A. (2020). Comparison of the CatBoost Classifier with other Machine Learning Methods. In Article in International Journal of Advanced Computer Science and Applications (Vol. 11, Issue 11). https://www.researchgate.net/publication/348277609

Price, M., & Dalgleish, J. (2010). Cyberbullying: Experiences, impacts and coping strategies as described by Australian young people. Youth Studies Australia, 29(2), 51–59.

Rasel, R. I., Sultana, N., Akhter, S., & Meesad, P. (2018). Detection of cyber-aggressive comments on social media networks: A machine learning and text mining approach. ACM International Conference Proceeding Series, September, 37–41. https://doi.org/10.1145/3278293.3278303

Reynolds, K., Kontostathis, A., & Edwards, L. (2011). Using machine learning to detect cyberbullying. Proceedings - 10th International Conference on Machine Learning and Applications, ICMLA 2011, 2, 241–244. https://doi.org/10.1109/ICMLA.2011.152

Rezvani, N., Beheshti, A., & Tabebordbar, A. (2020). Linking textual and contextual features for intelligent cyberbullying detection in social media. ACM International Conference Proceeding Series, 3–10. https://doi.org/10.1145/3428690.3429171

Talpur, B. A., & O’Sullivan, D. (2020). Cyberbullying severity detection: A machine learning approach. PLoS ONE, 15(10 October). https://doi.org/10.1371/journal.pone.0240924

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., & Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: an experimental review. Journal of Big Data, 7(1). https://doi.org/10.1186/s40537-020-00349-y

Van Hee, C., Jacobs, G., Emmery, C., DeSmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W., & Hoste, V. (2018a). Automatic detection of cyberbullying in social media text. PLoS ONE, 13(10), 1–22. https://doi.org/10.1371/journal.pone.0203794

Yang, F. J. (2018). An implementation of naive bayes classifier. Proceedings - 2018 International Conference on Computational Science and Computational Intelligence, CSCI 2018, 301–306. https://doi.org/10.1109/CSCI46756.2018.00065

Yang, L., Sun, L., Zhang, Z., Sun, Y., Wu, H., & Ye, D. (2014). Internet addiction, adolescent depression, and the mediating role of life events: Finding from a sample of Chinese adolescents. International Journal of Psychology, 49(5), 342–347. https://doi.org/10.1002/ijop.12063

Zhang, W., Bifet, A., Zhang, X., Weiss, J. C., & Nejdl, W. (2021). FARF: A Fair and Adaptive Random Forests Classifier. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12713 LNAI, 245–256. https://doi.org/10.1007/978-3-030-75765-6\_20

Zhao, R., Zhou, A., & Mao, K. (2016a). Automatic detection of cyberbullying on social networks based on bullying features. ACM International Conference Proceeding Series, 04-07-Janu, 1–6. https://doi.org/10.1145/2833312.2849567