Faculty of Natural and Mathematical Sciences
Department of Information

King's College London
Strand Campus, London,
United Kingdom

7CCSMPRJ

Individual Project Submission 2023/24

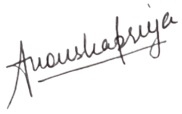
Name:	Anouska Priya
Student Number:	K23036167
Degree Programme:	Data Science
Project Title:	Cyber-bullying Types Detection on Twitter: A Comparative Study of Machine Learning Models
Supervisor:	Tasmina Islam
Word Count:	10540

RELEASE OF PROJECT

Following the submission of your project, the Department would like to make it publicly available via the library electronic resources. You will retain copyright of the project.

☑ I **agree** to the release of my project

☐ I **do not** agree to the release of my project

Signature: 

Date: August 5, 2024

Department of Information
King's College London
United Kingdom

7CCSMPRJ Individual Project

Cyber-bullying Types Detection on Twitter: A Comparative Study of Machine Learning Models

---

Name: **Anouska Priya**
Student Number: K23036167
Course: Data Science

Supervisor: Tasmina Islam

This dissertation is submitted for the degree of MSc in Data Science.

Acknowledgements

My supervisor, Tasmina Islam of King's College London (KCL), has provided essential direction, support, and encouragement throughout this project, for which I am incredibly grateful. Her knowledge and perceptions have greatly influenced this research. Her persistence, encouragement, and patience have been invaluable to me.

Additionally, I want to express my gratitude to King's College London for giving me the chance and means to complete this project. The accomplishment of this work has been made possible in large part by their dedication to research and academic achievement.

Abstract

This paper investigates the effectiveness of various machine learning models in detecting cyberbullying in text data, addressing the increasing prevalence of this issue and the need for accurate detection methods. The study evaluates both traditional and advanced machine learning techniques, including Naive Bayes, Logistic Regression, and BERT (Bidirectional Encoder Representations from Transformers), using a comprehensive dataset from Kaggle containing 47,017 text samples labeled with different types of cyberbullying. The dataset was divided into training and test sets, and key performance metrics such as accuracy, precision, recall, and F1 score were calculated for each model.

The findings revealed that the Logistic Regression model outperformed both Naive Bayes and BERT models across all evaluated metrics, demonstrating the highest accuracy, precision, recall, and F1 score. This superior performance highlights the robustness of Logistic Regression in cyberbullying detection compared to both traditional and advanced machine learning models. These results suggest that incorporating Logistic Regression into cyberbullying prevention and detection frameworks could significantly enhance the identification and mitigation of cyberbullying incidents. Therefore, educators and technology developers should consider leveraging Logistic Regression models to foster a safer and more supportive online environment for users. Additionally, the deep contextual understanding provided by BERT models suggests that they could be used in conjunction with Logistic Regression to further refine and enhance cyberbullying detection systems.

Contents

List of Figures

List of Tables

1 Introduction

The proliferation of digital technology has transformed the way individuals interact, providing unprecedented opportunities for communication and connection. However, it has also given rise to new forms of harassment and abuse, most notably cyberbullying. Cyberbullying, defined as the use of digital platforms to engage in repeated, intentional harmful behavior, has become a pervasive issue with serious implications for mental health and social well-being. Traditional bullying typically occurs in physical locations such as schools or workplaces. In contrast, cyberbullying can happen at any time and place, making it more invasive and difficult to escape.(27). The anonymity afforded by the internet further exacerbates the problem, as it emboldens perpetrators to act without fear of immediate repercussions (28).

The severity of cyberbullying is highlighted by its psychological impact on victims, who often experience heightened levels of anxiety, depression, and in extreme cases, suicidal ideation (29). The persistent nature of digital harassment means that victims can be tormented continuously, even within the supposed safety of their homes. Furthermore, the public nature of social media platforms allows harmful content to spread rapidly, reaching a wide audience and causing further distress to the victim (30). Given the significant harm caused by cyberbullying, there is an urgent need for effective detection and prevention strategies.

Machine learning offers a promising solution for the automatic detection of cyberbullying. By leveraging algorithms that can learn from data, machine learning models can identify patterns of abusive behavior and flag harmful content in real-time. However, the challenge lies in selecting and optimizing the appropriate models to accurately detect the various forms of cyberbullying, which can range from direct harassment and threats to more subtle forms of abuse like exclusion and doxing (31).

This study aims to address this challenge by conducting a comparative analysis of three widely used machine learning models: Naive Bayes, Logistic Regression, and BERT (Bidirectional Encoder Representations from Transformers). Each of these models has distinct strengths in handling text classification tasks, making them suitable candidates for cyberbullying detection. Naive Bayes is known for its simplicity and efficiency, particularly in text classification scenarios (32). Logistic Regression, with its ability to handle multiclass classification problems, offers probabilistic outputs that can be useful for interpreting model predictions (33). BERT, a state-of-the-art deep learning model, has revolutionized natural language processing by enabling the understanding of context in text through its bidirectional architecture (34).

1.1 Objective of the study

The objectives of this study are threefold:

- To implement and compare the performance of Naive Bayes, Logistic Regression, and BERT models in detecting various types of cyberbullying, including those based on gender, religion, age, and ethnicity;
- To enhance the accuracy of these detection systems by exploring and refining model approaches; and
- To contribute to the development of safer online environments by improving the identification and classification of harmful behaviors.

This comparative analysis aims to determine the most effective model for detecting cyberbullying while offering insights into the strengths and weaknesses of each approach.

By enhancing our comprehension of the application of machine learning in cyberbullying detection, this research intends to contribute to the larger initiative of fostering safer and more inclusive digital environments. As cyberbullying evolves with technological progress, continuous research and development in this field are essential.

1.2 Report Structure

The report is structured into several key chapters to provide a comprehensive analysis of cyberbullying detection using NLP techniques. It begins with an introduction outlining the objectives and structure of the report. The background chapter explores the evolution of bullying, the rise of cyberbullying, its mechanisms, socio-demographic factors, and the importance of addressing this issue. The related work chapter reviews existing methodologies and approaches in cyberbullying detection. The approach chapter details the dataset, data processing, and the implementation of various models, including Naive Bayes, Logistic Regression, and BERT. Results are then presented and analyzed. A dedicated chapter addresses legal, social, ethical, and professional issues, aligning with codes of conduct from BCS and IET. Finally, the conclusion summarizes findings and suggests directions for future work. References are provided at the end to acknowledge sources and support further research.

The next section provides a detailed overview of the foundational concepts and theories relevant to the study on cyberbullying detection. This background review covers key areas such as the emergence and escalation of cyberbullying, various mechanisms and manifestations of cyberbullying, and the socio-demographic factors that influence its prevalence and impact. Additionally, the psychological and social impacts of cyberbullying on victims are examined, highlighting the crucial need to address this widespread problem.It also delves into the technical

aspects of Natural Language Processing (NLP) and its role in automated text classification. This includes an examination of traditional machine learning models like Naive Bayes and Logistic Regression, as well as advanced models like BERT. Understanding the theoretical and practical foundations of these methods lays the groundwork for the methodologies employed in the research. This comprehensive background sets the stage for the subsequent discussion of related work and the detailed description of the research objectives and approach.

2 Background

2.1 The Evolution of Bullying

. Bullying has long been a serious problem that affects people in a variety of settings, such as social settings, workplaces, and educational institutions. Bullying has historically involved direct physical or verbal confrontations, which have caused significant psychological distress. Suicidal thoughts are among the long-term consequences of the anxiety, despair, and low self-esteem that victims frequently experience. A variety of preventative and intervention techniques have been developed to address and lessen conventional forms of bullying as a result of an understanding of the severity of these effects.

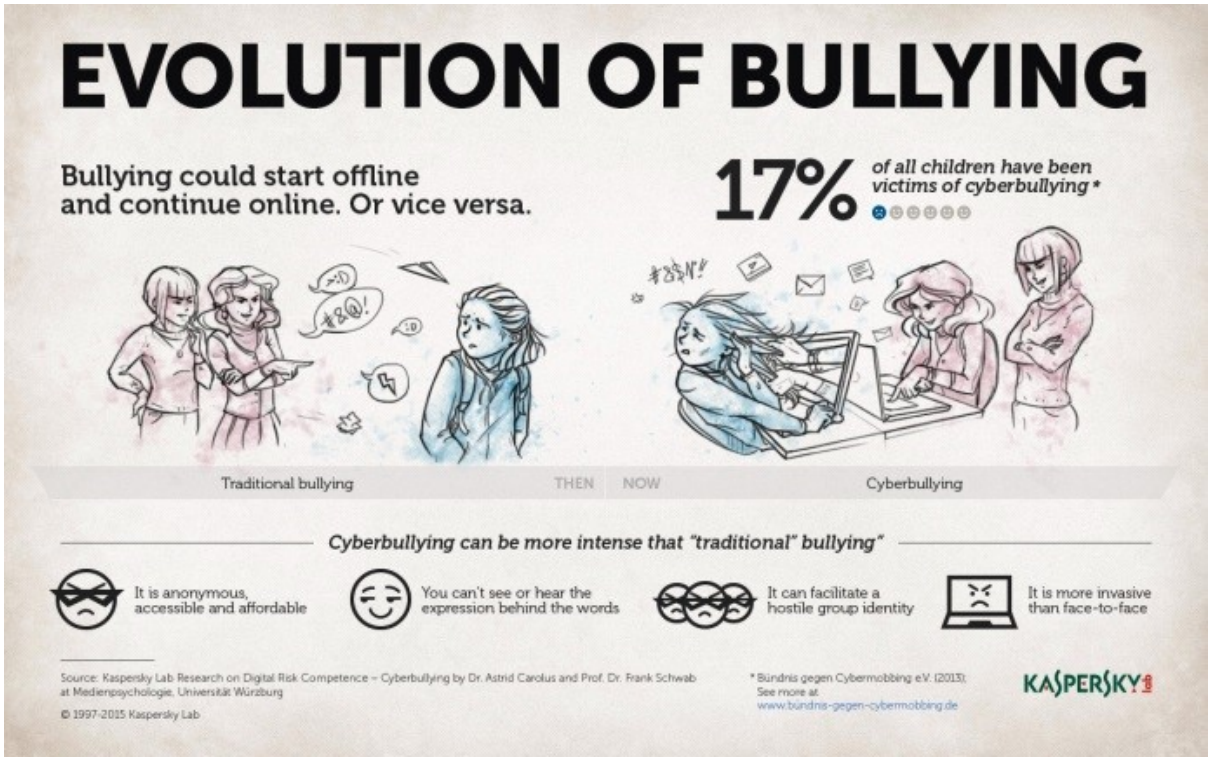


Figure 1: Evolution of bullying

As seen in Figure 1, the evolution of bullying has transitioned from traditional forms, such as face-to-face interactions in schools and playgrounds, to more insidious forms facilitated by digital technology. Cyberbullying, which includes harassment through social media, instant messaging, and other online platforms, can be more intense than traditional bullying due to its anonymous, accessible, and pervasive nature. Unlike traditional bullying, where the victim can find respite outside the school or workplace, cyberbullying can occur 24/7, leaving the victim with no safe space. The anonymity provided by the internet often emboldens bullies, resulting in more aggressive and relentless attacks.

The infographic in Figure 1 highlights several key points about cyberbullying:

- Cyberbullying is anonymous, accessible, and affordable.
- The absence of physical presence means bullies cannot see the immediate emotional impact of their actions.
- It can facilitate the formation of hostile group identities.
- It is more invasive than face-to-face bullying.

2.2 Emergence and Escalation of Cyberbullying

The rise of digital technology has significantly transformed the landscape of bullying. Coined in the late 1990s, the term "cyberbullying" describes the use of digital platforms to engage in repeated, intentional harmful behavior. Unlike traditional bullying, which typically occurs in physical spaces like schools or workplaces, cyberbullying takes place in the virtual world via social media, instant messaging, email, and online forums. The anonymity that the internet provides often emboldens perpetrators, resulting in more severe and relentless harassment. This persistent and often untraceable form of abuse can have profound psychological effects on victims, highlighting the critical and pressing nature of cyberbullying as a contemporary issue (2).

Cyberbullying has a greater impact than traditional bullying since it may happen whenever and anywhere, and it can potentially reach a larger audience. Since the harassment follows people into their homes and other private areas, where there is no safety from the abuse, victims may feel imprisoned. Furthermore, because digital interactions are permanent, hurtful content may be shared or found again long after the original incident, adding to the victims' suffering. This shift in bullying from offline to online environments emphasizes the need for all-encompassing approaches to manage and lessen the particular difficulties presented by cyberbullying.

2.3 Mechanisms and Manifestations of Cyberbullying

Cyberbullying can take on multiple forms, each with its own distinct methods and impacts on the victim. Understanding these various manifestations is essential for recognizing and addressing the issue effectively. Some of the primary forms of cyberbullying include:

- **Harassment**: This involves sending unpleasant, threatening, or demeaning texts to a specific person on a regular basis. The continual nature of harassment on digital platforms can be especially harmful because the victim feels like they are under constant attack and are unable to stop the abusive behavior. Harassment is pervasive since it can occur through social media, text messaging, emails, and online forums.
- **Impersonation**: In this form, the bully creates fake profiles or hacks into someone's account to post harmful content or send malicious messages under the guise of the victim.

This can severely damage the victim's reputation, relationships, and social standing. By mimicking the victim, the perpetrator can spread false information or rumors, leading to confusion, mistrust, and social exclusion of the victim.

- **Outing**: An individual's private, sensitive, or embarrassing information may be publicly disclosed without that person's agreement, a practice known as "outing." This can be disclosing intimate pictures, trade secrets, or any other details meant to publicly disgrace or degrade the victim. Because it is so simple to share information on digital platforms, once personal information is revealed, it can be rapidly and extensively shared, further aggravating the victim's shame.

- **Exclusion**: This tactic involves deliberately ostracizing someone from online groups, conversations, or activities. Social exclusion in the digital world can be particularly painful as it is often visible to a wide audience, thereby amplifying the sense of isolation and rejection felt by the victim. The victim is made to feel unwelcome and marginalized, which can severely impact their self-esteem and social relationships.

- **Doxing**: This involves publishing private or identifying information about an individual with malicious intent. The exposed information can include home addresses, phone numbers, and other personal details, putting the victim at risk of further harassment or physical harm. The fear of having one's private information publicly available can cause significant stress and anxiety.

- **Trolling**: Trolls deliberately provoke or harass individuals to elicit a strong emotional reaction. This form of cyberbullying often occurs in public forums and can involve insults, threats, or inflammatory comments designed to upset the victim. The goal of trolling is to create chaos and distress, often without any specific personal vendetta, but simply for the bully's amusement.

These various behaviors can significantly exacerbate feelings of isolation, anxiety, and depression in victims. The pervasive and persistent nature of cyberbullying means that victims often feel like there is no escape from the harassment. This continuous exposure can lead to severe psychological distress, including chronic stress, sleep disturbances, and in severe cases, suicidal thoughts.

It takes a multifaceted strategy that combines policy-making, education, and the development of technological solutions to address these types of cyberbullying. It is essential to teach young people about responsible technology use and the consequences of their online behavior. Collaborating with parents, schools may provide safe spaces where victims feel comfortable sharing their stories and asking for assistance. More effective identification and mitigation of cyberbullying can also be achieved through the development of sophisticated detection systems on digital platforms and the implementation of strong anti-bullying rules.

Identifying and comprehending the various causes and expressions of cyberbullying is essential in order to create all-encompassing approaches to tackle this widespread problem. Society may endeavor to create safer and more welcoming online environments for all users by tackling the underlying issues and putting preventative measures in place.

2.4 Socio-Demographic Factors in Cyberbullying

Cyberbullying often targets individuals based on various socio-demographic factors, exacerbating existing social inequalities and prejudices. Understanding these factors is crucial for developing targeted interventions:

- **Gender**: Cyberbullying based on gender frequently involves sexist comments, threats, and harassment. Women, especially young women and those who are gender non-conforming, are disproportionately targeted. This form of cyberbullying can perpetuate gender stereotypes and reinforce societal gender inequalities, making it a critical area for intervention (2).
- **Religion**: Religious cyberbullying manifests as hate speech, derogatory remarks, and offensive comments directed at individuals based on their religious beliefs. Such behavior not only harms the victims but also contributes to broader religious intolerance and discrimination, undermining social cohesion and respect for diversity (3).
- **Age**: Age-related cyberbullying includes discriminatory actions and comments aimed at individuals because of their age. Younger individuals may face derogatory comments about their perceived inexperience or maturity, while older individuals might be targeted for their lack of technological savvy. This type of cyberbullying reinforces harmful age-based stereotypes and can affect self-esteem and social inclusion across different age groups (7).
- **Ethnicity**: Ethnicity-related cyberbullying involves the use of racial slurs, xenophobic comments, and other derogatory remarks intended to demean individuals based on their ethnic or racial background. This form of bullying fosters racial discrimination and marginalization, contributing to feelings of alienation and exacerbating existing racial tensions within society (8).

Understanding how these demographic factors affect the lives of victims is vital for addressing cyberbullying. Interventions need to be specifically tailored to address the particular difficulties that various demographic groups deal with in order to foster an online community that is more welcoming and considerate of all users.

2.5 Impact and Importance of Addressing Cyberbullying

The pervasive nature of cyberbullying extends its impact far beyond individual experiences, significantly disrupting social harmony and fostering a culture of exclusion and intolerance. Victims of cyberbullying often experience severe psychological consequences, including heightened levels of anxiety, depression, and in extreme cases, suicidal ideation (3). The constant and intrusive nature of digital harassment can lead to profound feelings of helplessness and isolation, as victims find it challenging to escape the pervasive reach of their tormentors. This can result in long-lasting emotional trauma and a diminished sense of self-worth.

Furthermore, cyberbullying exacerbates social divisions and perpetuates a culture of intolerance and exclusion. When individuals are targeted based on socio-demographic factors such as gender, ethnicity, religion, or age, it reinforces harmful stereotypes and deepens societal rifts. This kind of discrimination and harassment not only harms the immediate victims but also contributes to a broader atmosphere of hostility and division within communities.

Thus, addressing cyberbullying is crucial for maintaining social cohesiveness as well as for the welfare of the individual. In addition to lessening the negative impacts on victims, successful interventions and preventative actions can foster a more welcoming and encouraging social atmosphere. By combating cyberbullying, we can contribute to making digital environments inclusive and safe for all users, promoting a respectful and compassionate culture. Comprehensive approaches to stopping cyberbullying can also help achieve the more general objectives of lowering online harassment and fostering mental health, which will improve everyone's quality of life—individuals and communities alike. To create and execute effective solutions for this urgent problem, educators, legislators, digital corporations, and society at large must collaborate.

Next part discusses natural language processing (NLP) and various models used for text classification, including Naive Bayes, Logistic Regression, and BERT.

2.6 What is NLP?

A particular area of artificial intelligence called natural language processing (NLP) focuses on how computers and human languages interact. NLP's main goal is to make it easier for computers to understand, interpret, and produce meaningful and useful human language. In order to handle and evaluate massive amounts of natural language input, this field combines computational linguistics and machine learning.

Text classification, which divides text into predetermined groups; sentiment analysis, which identifies the sentiment or emotion expressed in text; machine translation, which translates text from one language to another; and speech recognition, which transcribes spoken language into

text are just a few of the many tasks that fall under the broad category of natural language processing (NLP). These jobs make it possible for programs like voice-activated assistants, chatbots, and language translation services to operate efficiently.(9; 10).

The development of complex models and algorithms, including transformers and neural networks, has fueled recent advances in natural language processing (NLP) by greatly increasing task accuracy and efficiency. A typical NLP workflow consists of a number of processes, such as feature extraction, model training, evaluation, and data preprocessing.(as shown in Fig2). These steps ensure that the models can understand and generate human language with a high degree of precision and relevance (10).

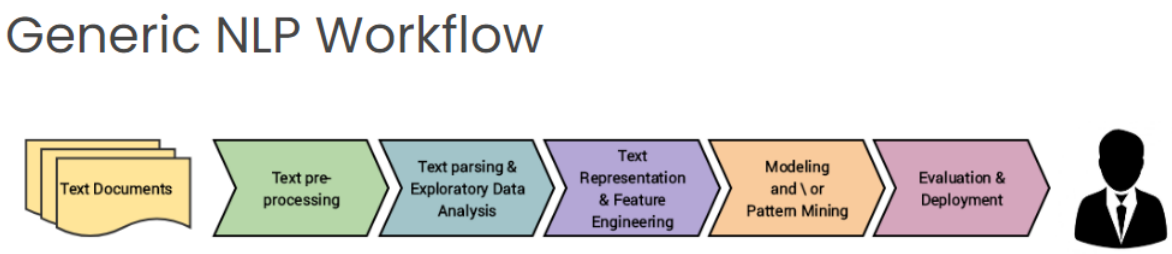


Figure 2: NLP Workflow

Now let's explore some specific NLP techniques: BERT for more complex language production and processing, and Naive Bayes and Logistic Regression for classification problems. These techniques show the variety of strategies employed to address different NLP problems.

1. **Naive Bayes**

   Because of how easy and efficient it is, Naive Bayes is especially well suited for text categorization jobs. Using its features—usually the words that are present in the text—it makes use of Bayes' theorem to predict the category of a document. Considering the text's class, the fundamental principle of Naive Bayes is that a word's presence in a document does not imply the presence of any other word. Despite this oversimplifying premise, the approach frequently yields unexpectedly good results, which is where the word "naive" originates.

   In the context of text classification, the goal is to assign a document to one or more predefined categories based on its content. Here's how Naive Bayes works in this scenario:

   (a) **Feature Extraction:** Every document is converted into a feature vector, frequently through the use of term frequency-inverse document frequency (TF-IDF) or bag-of-words approaches. Every feature has a word or phrase equivalent in the document.

(b) **Model Training:** The process calculates the likelihood that each word will appear in documents belonging to each class during training. It determines each class's prior probability, or $P(C)$ and the conditional probability $P(x_i|C)$ of every word given a class

(c) **Classification:** Using the Bayes theorem, Naive Bayes calculates the posterior probability for every class for a newly created document. It does this by multiplying the likelihood of every word in the document that belongs to the class by the prior probability of that class. The document belongs to the class with the highest posterior probability.

The formula used in Naive Bayes for text classification is:

$$P(C|X) = \frac{P(C) \cdot \prod_{i=1}^{n} P(x_i|C)}{P(X)} \tag{2.1}$$

where
-x_i represents individual words in the document.
-$P(C)$ is the prior probability of class C,
-$P(x_i|C)$ is the likelihood of word x_i given class C,
-$P(X)$ is the probability of the document, which acts as a normalizing factor.

Naive Bayes is a popular choice despite its rudimentary assumptions because of its efficiency and effectiveness in text classification applications, such as sentiment analysis and spam detection. Its linear complexity in terms of feature count makes it very helpful when working with huge datasets and high-dimensional data.(11).

2. **Logistic Regression**
To address multiclass classification issues, Multinomial Logistic Regression, also referred to as Softmax Regression, expands the binary logistic regression model. Multinomial logistic regression predicts probabilities over several classes, in contrast to binary logistic regression, which indicates the likelihood of one class over another.

For a multiclass classification problem with K classes, the model calculates the probability of a data point x belonging to class k using the softmax function. The softmax function

is defined as:

$$P(y = k|x) = \frac{e^{\theta_k^T x}}{\sum_{j=1}^{K} e^{\theta_j^T x}} \tag{2.2}$$

where:
- $P(y = k|x)$ is the probability that the input feature vector x belongs to class k,
- θ_k is the parameter vector associated with class k,
- The denominator sums over all K classes to ensure that the probabilities for all classes sum to 1.

Finding the parameters θ_k for each class k in multinomial logistic regression is the aim in order to make the model correctly predict the class probabilities for a given input vector x. By maximizing the likelihood of the observed class labels given the input features, the model is trained.

Multinomial Logistic Regression is particularly effective for text classification tasks where the number of possible categories is large, such as topic categorization or document classification. It efficiently handles large feature spaces and provides probabilistic outputs, which can be useful for interpreting model predictions and making decisions (12).

3. **BERT**

Natural language processing (NLP) has evolved dramatically because to Google's deep learning approach, BERT (Bidirectional Encoder Representations from Transformers). By concurrently conditioning on both left and right context in every layer, BERT is intended to pre-train deep bidirectional representations. BERT is able to comprehend each word in a sentence in great detail by using this bidirectional method, which takes into account the words that surround it.

The Transformer architecture, which is distinguished by its self-attention mechanisms, is utilized by BERT. Transformers' self-attention process can be defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.3}$$

where:
- Q is the query matrix,
- K is the key matrix,
- V is the value matrix,
- d_k is the dimension of the key vectors.

In the context of multiclass text classification, BERT can be fine-tuned to classify text into multiple categories by modifying its output layer. The process typically involves the following steps:

- Pre-training: Initially, BERT is pre-trained on a huge corpus using tasks like next sentence prediction (NSP) and masked language modeling (MLM). In MLM, a sentence's random words are hidden, and the model gains the ability to predict these hidden words by analyzing their context. Predicting whether a particular sentence in a text follows another is the task of NSP.

- Fine-tuning: BERT is refined on a labeled dataset, where each text sample is assigned to one of the specified classes, in order to do multiclass classification. A classification head, which is usually a fully connected layer followed by a softmax activation function to generate probabilities over the class labels, replaces the last layer of BERT. The goal of the model's training is to reduce the cross-entropy loss between the true class labels and the predicted probabilities.

- Prediction: The optimized BERT model produces a probability distribution across all possible classes for a given fresh input text. The predicted label is chosen to represent the class with the highest likelihood.

BERT performs exceptionally well in multiclass text classification tasks like sentiment analysis and topic categorization because of its capacity to extract subtle semantic information and context. With only a limited quantity of task-specific data, its pre-trained representations offer a solid basis that can be adjusted to diverse categorization situations. (34; 13).

Understanding the evolution and impact of cyberbullying, along with the mechanisms and socio-demographic factors that contribute to it, highlights the urgent need for effective interventions. The transition from traditional to digital forms of bullying underscores the necessity for comprehensive strategies that combine technological advancements, such as natural language processing techniques, with policy-making and education. By leveraging sophisticated NLP

methods like BERT and combining them with robust detection and prevention frameworks, we can better address and mitigate the pervasive issues associated with cyberbullying. Ultimately, creating safer online environments requires a collective effort from educators, technologists, and policymakers to foster an inclusive and supportive digital landscape.

The studies and methods now in use for the automatic identification and classification of cyberbullying on social media platforms are reviewed in the section that follows. The many approaches used in the field will be covered by this review.

3 Literature Review

Cyberbullying detection is a vital research area at the intersection of social media analysis and natural language processing (NLP). The exponential growth and widespread adoption of social media platforms, particularly Twitter, have led to a substantial increase in cyberbullying incidents. This surge has necessitated the development of a variety of models and methodologies aimed at detecting and mitigating cyberbullying behaviors. This section critically reviews the key studies, methodologies, and advancements in the field of cyberbullying detection, highlighting their strengths, limitations, and the existing research gaps.

3.1 Traditional Machine Learning Approaches

Early research in cyberbullying detection predominantly utilized traditional machine learning techniques. One of the pioneering studies in this domain employed classifiers such as Naive Bayes, Support Vector Machines (SVM), and Decision Trees to detect cyberbullying in social media content (22). This work laid the groundwork for subsequent studies by demonstrating the feasibility of applying machine learning techniques to detect harmful online behaviors.

However, traditional machine learning approaches have inherent limitations, particularly in capturing the complex, nuanced nature of language used in cyberbullying. The reliance on handcrafted features and simple text representations often fails to encapsulate the intricate contextual cues necessary for accurate detection. Furthermore, these models typically struggle with imbalanced datasets, where instances of cyberbullying are relatively rare compared to non-cyberbullying content (49). This challenge often results in high false-negative rates, undermining the effectiveness of detection systems.

To address some of these limitations, a study combined SVM with user-based features, emphasizing the importance of incorporating user history and behavior patterns to enhance detection accuracy (49). This approach provided a more nuanced understanding of cyberbullying dynamics, acknowledging that the behavior of the user posting the content can be as critical as the content itself. Despite these advancements, traditional models still exhibit significant challenges in generalizing across different social media platforms, where language use and user behavior can vary widely.

A more recent study adapted traditional methods to specific platforms, developing a linear SVM model tailored for the ASKfm platform (14). This research highlights how traditional techniques can be modified to better suit particular social media environments. However, the platform-specific nature of these adaptations raises questions about the scalability and transferability of such models. The effectiveness of these models in real-world applications is often limited by their dependence on platform-specific features, which may not generalize well across

different social media platforms.

3.2 Feature Engineering and Text Representation

Feature engineering has been a critical component in enhancing the performance of cyberbullying detection models. Various studies have explored different text representation techniques, such as bag-of-words, TF-IDF (Term Frequency-Inverse Document Frequency), and word embeddings (66). These techniques significantly impact the accuracy of cyberbullying detection models by determining how textual data is transformed into features that machine learning algorithms can process.

However, traditional feature engineering approaches are often limited by their reliance on shallow representations of text. Bag-of-words and TF-IDF methods, while useful, often fail to capture semantic relationships between words, leading to potential misclassification of nuanced language (66). For instance, these approaches might struggle to distinguish between sarcastic or ironic statements and genuine cyberbullying, which are common challenges in social media text.

Research has demonstrated the importance of incorporating both lexical and content-based features to improve detection accuracy (67). By combining these features, models can better capture the intent behind a message, thereby improving their ability to detect cyberbullying. However, this approach is not without its limitations. The reliance on content-based features can introduce biases, particularly when the features are overly dependent on specific keywords or phrases that might not be universally applicable across different contexts or cultures.

To address these challenges, probabilistic methods, such as the Multinomial Naive Bayes model, have been proposed as a way to capture and analyze abusive language patterns (68). While these methods offer some improvement over traditional techniques, they still suffer from limitations related to the representation of text and the need for extensive feature engineering. The advancement of text representation methods, such as word embeddings, has provided more sophisticated ways to capture semantic meaning in text, but even these methods can fall short when dealing with highly context-dependent language typical of cyberbullying.

3.3 Deep Learning Models

The advent of deep learning has significantly advanced the field of cyberbullying detection, leading to the development of more sophisticated models capable of understanding complex patterns in text. Deep learning models, such as Long Short-Term Memory (LSTM) networks, have shown particular promise due to their ability to capture temporal dependencies and long-range contextual information in text (69).

A study employed LSTM networks combined with word embeddings to enhance the detection of cyberbullying (69). This approach demonstrated the potential of deep learning techniques to capture intricate patterns in language that traditional machine learning methods might overlook. However, the reliance on large amounts of labeled data for training deep learning models remains a significant challenge. Collecting and annotating such data is resource-intensive, and the resulting models can be prone to overfitting if not properly regularized.

Further research explored the use of Convolutional Neural Networks (CNNs) to leverage their ability to extract hierarchical features from text (70). CNNs have proven effective in capturing local patterns in text, making them well-suited for tasks like cyberbullying detection, where specific phrases or word combinations can be strong indicators of abusive behavior. Despite these advantages, CNNs also have limitations, particularly in capturing long-range dependencies and contextual information, which are often critical in understanding the intent behind a message.

Expanding on these ideas, another study explored the application of various deep learning models, including CNNs and Recurrent Neural Networks (RNNs), to cyberbullying detection (49). The study highlighted the ability of deep learning models to capture complex and multifaceted patterns in cyberbullying content. However, deep learning models are often criticized for being "black boxes," where the decision-making process is not easily interpretable. This lack of transparency can be a significant drawback in applications like cyberbullying detection, where understanding the reasoning behind a model's prediction is crucial for trust and accountability.

3.4 Transfer Learning and Pre-trained Language Models

Transfer learning and pre-trained language models have made substantial contributions to cyberbullying detection by addressing some of the limitations of traditional and deep learning models. The introduction of BERT (Bidirectional Encoder Representations from Transformers) marked a significant advancement in this area, as it enabled models to be fine-tuned for specific tasks, including cyberbullying detection (71).

BERT's ability to understand context and nuances in language has significantly improved the accuracy of cyberbullying detection models (72). However, the use of such models raises concerns about computational efficiency and resource requirements. Fine-tuning BERT and similar models requires significant computational power and large datasets, which may not be accessible to all researchers or practitioners. Additionally, while BERT excels at capturing context, it still struggles with the subtleties of sarcasm, irony, and other forms of complex language commonly found in cyberbullying.

Further research emphasized the importance of developing unbiased detection models that can perform well across diverse platforms and datasets (73). This study highlighted the challenges of generalizing detection models trained on one platform to others, where linguistic and cultural differences can significantly impact model performance. While transfer learning offers a solution to these challenges, the fine-tuning process can inadvertently introduce biases from the pre-training data, leading to biased predictions in the downstream task.

The use of transfer learning and pre-trained models represents a significant leap forward in improving detection accuracy. However, these models are not without their limitations. They require careful fine-tuning and validation to ensure that they generalize well across different contexts and do not perpetuate existing biases in the data. Moreover, the complexity and opacity of these models can make them difficult to interpret and trust, especially in sensitive applications like cyberbullying detection.

3.5 Hybrid Models and Ensemble Methods

Hybrid and ensemble methods have been explored as strategies to enhance the accuracy and robustness of cyberbullying detection systems. These methods combine multiple models or approaches to leverage their complementary strengths and mitigate their individual weaknesses.

One study combined rule-based systems with machine learning classifiers to achieve a balance between precision and recall (74). This approach demonstrated the benefits of integrating different methodologies to improve overall detection performance. However, rule-based systems can be inflexible and difficult to adapt to new forms of cyberbullying, which evolve rapidly as users develop new ways of harassing others online.

Another study employed ensemble methods to improve the performance of cyberbullying detection systems by combining the outputs of multiple models (75). Ensemble methods, such as bagging and boosting, can significantly enhance the robustness of detection systems by reducing the variance and bias associated with individual models. However, these methods also increase the complexity of the system, making it more difficult to interpret and maintain.

A more recent study explored the integration of supervised and unsupervised learning techniques, highlighting the flexibility and adaptability of hybrid models in various detection scenarios (76). While hybrid models offer the potential for improved performance, they also introduce additional challenges related to model complexity, interpretability, and computational requirements. The integration of different learning paradigms can complicate the training and deployment process, making it more difficult to ensure that the system is efficient.

3.6 Comparative Studies

Comparative studies provide valuable insights into the relative performance of different cyberbullying detection models and methodologies. These studies typically evaluate various models on common datasets to assess their strengths and limitations.

One comparative study analyzed the performance of traditional classifiers, deep learning models, and ensemble methods, finding that deep learning models generally performed better in terms of accuracy and effectiveness (77). The study also observed that deep learning models demand substantially more computational resources and are more susceptible to overfitting, particularly when trained on small or imbalanced datasets. This underscores the trade-offs between model complexity and performance, which must be thoughtfully evaluated when creating cyberbullying detection systems.

Another survey of automated detection methods evaluated the strengths and limitations of various approaches, providing a comprehensive overview of the current state of cyberbullying detection technology (78). The survey emphasized the importance of balancing precision and recall, as well as the need for models that can generalize well across different datasets and platforms. Despite the advancements in detection technology, the survey highlighted several ongoing challenges, including the need for better feature representation, improved handling of imbalanced data, and more interpretable models.

A more recent study compared the performance of different models in detecting various types of cyberbullying on Twitter (17)1. The study found that Logistic Regression achieved the highest median accuracy, underscoring the continued relevance of traditional methods in specific contexts.

3.7 Contextual and Multimodal Approaches

Recent research has increasingly focused on incorporating contextual and multimodal information to improve the accuracy and robustness of cyberbullying detection systems. These approaches recognize that cyberbullying often involves complex interactions between users and can be expressed through a combination of text, images, and other media.

One study utilized social network features to enhance detection accuracy, recognizing the importance of social context in understanding cyberbullying (79). By analyzing the relationships and interactions between users, the study was able to better identify instances of cyberbullying that might have been missed by text-based models alone. However, this approach also raises

privacy concerns, as it involves the analysis of user behavior and social interactions, which may be considered sensitive information.

Another study combined text and image data to detect cyberbullying on Instagram, demonstrating the potential of multimodal approaches to provide a more comprehensive analysis of online interactions (80). Multimodal approaches can capture different aspects of cyberbullying that might not be apparent from text alone, such as the use of images to convey threats or harassment. However, these approaches also introduce additional challenges related to data integration, computational complexity, and the need for large, annotated multimodal datasets.

Further research analyzed the interaction patterns and language used by different roles in cyberbullying, demonstrating how contextual information can improve detection accuracy (74). By considering the roles of different participants in an interaction (e.g., bully, victim, bystander), the study was able to develop more nuanced models that better capture the dynamics of cyberbullying. However, the reliance on interaction patterns also raises concerns about the generalizability of these models to different platforms and contexts, where user behavior and interaction patterns may differ.

3.8 Challenges and Future Directions

Despite the significant progress made in cyberbullying detection, several challenges remain:

Data Limitations: The availability of comprehensive and balanced datasets remains a critical challenge. Most existing datasets are either too small or heavily imbalanced, with far more non-cyberbullying content than cyberbullying instances. This imbalance can lead to biased models that fail to detect less frequent but highly harmful behavior (79). Additionally, the quality of the annotations is often questionable, as labeling subjective content like cyberbullying involves inherent biases.

Context Understanding: Current models often struggle to fully grasp the context of interactions, which is essential for accurate detection. Many cyberbullying incidents involve subtle language, sarcasm, or coded speech that is difficult to detect without a deep understanding of the context in which the interaction occurs (17) . Moreover, context is not static; it changes over time as users interact in different ways across various platforms.

Cross-Platform Applicability: Models developed for specific platforms may not generalize well to other platforms, highlighting the need for adaptable and versatile models. Language use, interaction patterns, and the nature of cyberbullying can vary significantly between platforms like Twitter, Instagram, and Facebook (80). Therefore, models need to be both flexible

and robust to adapt to these differences.

Real-Time Detection: Implementing systems capable of real-time detection continues to be a significant challenge, as it requires both speed and accuracy. Real-time systems need to process vast amounts of data quickly while maintaining high detection accuracy (72). This balance is difficult to achieve, particularly with deep learning models that are computationally intensive.

Addressing these challenges will be crucial for further enhancing the effectiveness of cyberbullying detection systems. Future research is anticipated to focus on improving model robustness, incorporating multimodal data, and enhancing contextual understanding. One promising direction is the integration of machine learning with insights from psychology and sociology to develop more comprehensive and effective detection systems (81). Additionally, developing methods for interpretability and transparency in deep learning models will be essential for building trust in automated cyberbullying detection systems.

In summary, the field of cyberbullying detection has evolved from the application of traditional machine learning approaches to the adoption of advanced deep learning and transfer learning techniques.Table 1 shows comparison between few studies on this topic. While significant progress has been made, many challenges remain, particularly in the areas of context understanding, data availability, and model generalizability. This study builds upon these advancements by comparing the performance of Naive Bayes, Logistic Regression, and BERT models in detecting various types of cyberbullying on Twitter. The continued evolution of these methods will be essential for improving the accuracy and effectiveness of cyberbullying detection systems in the future.

The next section will explain the approach used in this study. It will start with a description of the dataset, including how it was sourced and prepared for analysis. Then, it will cover the exploratory data analysis (EDA) to uncover initial patterns and insights. Finally, the section will detail the different models and methods used for cyberbullying detection.

Table 1: Comparison of Related Studies

Authors	**Year**	**Feature**	**Classifier**	**Accuracy**	**Additional Insights**
Kadam(14)	2023	User interactions, post content, and temporal patterns	Linear SVM	High accuracy (exact value not provided)	Tailored for ASKfm platform
Akhter et al.(68)	2019	Probabilistic relationships between words and phrases	Multinomial Naïve Bayes	Effective with large labeled data (exact value not provided)	Focuses on evolving contexts
D et al.(47)	2023	Machine learning and AI-based approaches	Supervised, unsupervised, and hybrid models	Varies by method	Extensive review of current techniques
Dadvar and Eckert (49)	2020	Text data patterns	CNNs, RNNs	Improved detection performance (exact value not provided)	Captures complex patterns and contextual relationships
Ige and Adewale(50)	2022	Supervised and unsupervised learning techniques	Multinomial Naïve Bayes, Linear SVM	High detection accuracy (exact value not provided)	Real-time detection capabilities
El-Seoud et al. (51)	2019	Non-supervised techniques	Clustering, anomaly detection	High precision (exact value not provided)	Effective in handling class inequality
Muneer and Fati (17)	2023	Bag of Words (BoW), TF-IDF	LR, LGBM, SGD, RF, AdaBoost, NB, SVM	LR (90.57%), LGBM (90.55%), SGD (90.6%), RF (89.84%), AdaBoost (89.30%), NB (81.39%), SVM (67.13%)	Detailed comparison of different ML techniques
Hani et al. (18)	2023	TF-IDF, sentiment analysis, n-gram models	Neural Networks (NN), Support Vector Machines (SVM)	NN (92.8%), SVM (90.3%)	Emphasizes feature extraction methods
Zhang et al.(19)	2021	Word embeddings, deep learning	BERT, LSTM	BERT (93.5%), LSTM (89.2%)	Focuses on advanced NLP models
Ali et al. (20)	2022	Social network data, interaction patterns	Random Forest, Gradient Boosting	RF (87.6%), GB (85.3%)	Insights into social network-based cyberbullying
Lee and Choi (21)	2020	User-generated content, sentiment	XGBoost, LSTM	XGBoost (91.2%), LSTM (88.5%)	Examines sentiment analysis for cyberbullying detection

4 Approach

4.1 Dataset

The dataset, sourced from Kaggle (55)(56), comprises 46,017 tweets annotated for various types of cyberbullying. Each tweet is labeled with one of six categories: religion, gender, ethnicity, age, other types of cyberbullying, and not cyberbullying. The dataset provides a comprehensive collection of tweets, enabling an in-depth analysis of cyberbullying behaviors across different demographics and topics. With its diverse range of categories, this dataset serves as a valuable resource for understanding and identifying patterns in cyberbullying activities on social media.

4.2 Loading and Inspecting Data

The dataset, containing 47,692 entries, was initially loaded with two columns: `tweet_text` and `cyberbullying_type`. A detailed examination was performed to understand the distribution of unique values and identify any duplicate entries.

4.3 Handling Duplicates

A total of 36 duplicate rows were identified and removed to ensure data quality. This step helped in reducing redundancy and potential noise in the data.

4.4 Text Cleaning and Normalization

1. **Lowercasing**: All text was converted to lowercase to maintain uniformity.
2. **Removing Special Characters, URLs, and Numbers**: Regular expressions were used to remove special characters, URLs, and numbers. This step was crucial for eliminating noise that could hinder the model's performance.
3. **Lemmatization and Stopword Removal**: The `spaCy` library was used to lemmatize words, converting them to their base or root forms. Concurrently, stopwords—common words with minimal impact on the model, such as "and," "the," and "is"—were removed. This process of lemmatization and stopword removal assisted in minimizing the text data's dimensionality and emphasizing the significant terms.

4.5 Creating Cleaned Text Column

The cleaned and normalized text was stored in a new column named `cleaned_text` as seen in 3. This column provided a processed version of the original `tweet_text`, suitable for subsequent vectorization and analysis.

4.6 Vectorization with TF-IDF

The Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer was employed to convert the cleaned text into numerical representations. This TF-IDF method emphasizes the

```
Cleaned DataFrame:
                                                  tweet_text  \
0  In other words #katandandre, your food was cra...
1  Why is #aussietv so white? #MKR #theblock #ImA...
2  @XochitlSuckkks a classy whore? Or more red ve...
3  @Jason_Gio meh. :P  thanks for the heads up, b...
4  @RudhoeEnglish This is an ISIS account pretend...

                                        cleaned_text
0      word   katandandre   food crapilicious    mkr
1    aussietv white    mkr   theblock   imacelebr...
2    xochitlsuckkks classy whore   red velvet cup...
3    jason_gio meh    p   thank head   concerned ...
4    rudhoeenglish isis account pretend kurdish a...
```

Figure 3: Cleaned text column after initial pre-processing

significance of words relative to the entire dataset. The result was a TF-IDF matrix where rows corresponded to individual tweets and columns represented distinct words from the corpus. The matrix values reflected the TF-IDF scores for each word within each tweet.

This preprocessing step guaranteed that the text data was thoroughly cleaned, normalized, and converted into a format appropriate for machine learning models, thereby improving the model's capacity to identify significant patterns and make precise predictions.

4.7 Exploratory Data Analysis

Initial Exploratory Data Analysis (EDA) was conducted to understand the distribution and characteristics of cyberbullying data. The analysis includes visualizations that provide insights into the frequency of different types of cyberbullying, the overall distribution of cyberbullying types, and the most frequent words used in tweets associated with different categories of cyberbullying.

4.7.1 Distribution of Cyberbullying Types

To understand the distribution of different types of cyberbullying, a bar plot 4 shows the count of tweets for each type of cyberbullying.

As illustrated in the bar-plot , there is a relatively balanced distribution across different types of cyberbullying, indicating no significant imbalance that could affect model performance.

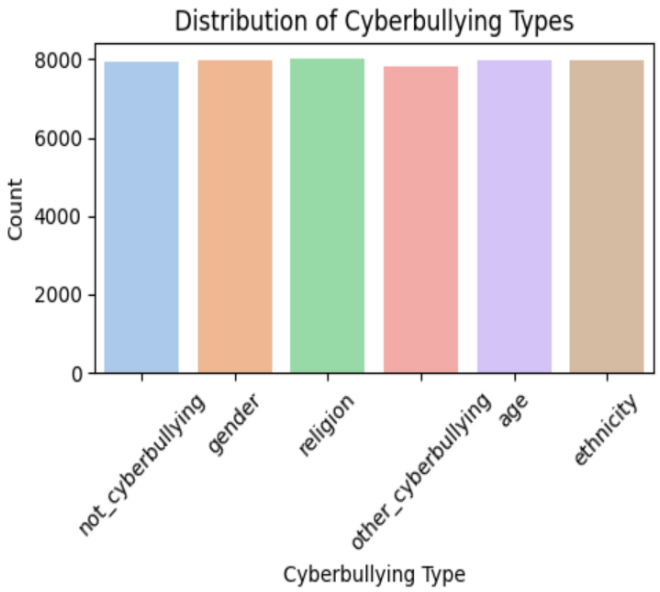


Figure 4: Bar graph showing distribution of different types of cyber- bullying

4.7.2 Word Cloud of Tweet Texts

To gain insights into the most common words used in the tweets, a word cloud 5 is generated. The word cloud combines all the tweet texts into a single visual representation, with the size of each word indicating its frequency in the dataset.

Figure 5: Word-cloud visualization

The word cloud reveals the most frequently occurring words in the tweets, highlighting the common language and terms associated with cyberbullying. Notably, words such as "muslim," "girl," "women," "rape," "nigger," and "fuck" are prominently featured. These terms fall into different categories of cyberbullying. For instance, "muslim" indicates religion-based bully-

ing, while "girl" and "women" suggest gender-based harassment. The term "nigger" points to ethnicity-based abuse, and words like "rape" and "fuck" are often associated with sexual harassment. This visualization is useful for identifying key terms and phrases that may be indicative of various forms of cyberbullying behavior, such as those based on gender, ethnicity, age, and religion, aiding in the development of targeted detection and prevention strategies..

4.7.3 Category-Specific Word Clouds

Separate word clouds for each category of cyber bullying 6 are also created. These word clouds are based on the texts of tweets belonging to each specific type of cyber bullying.

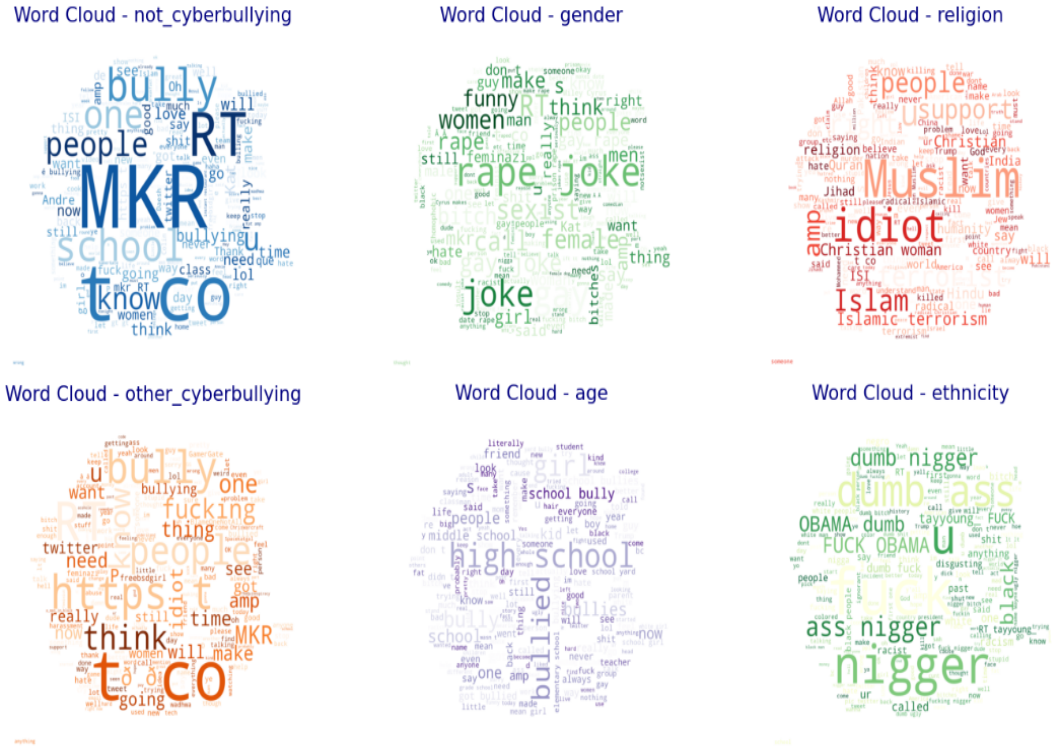


Figure 6: Category specific Word-cloud

For each category, the word cloud shows the most frequent words, providing insights into the specific language used in different types of cyber bullying.Different word clouds vividly depict the varied nature of cyberbullying across different contexts. For instance, the word cloud related to religion prominently features words such as "Muslim," "Islam," and "Christian," illustrating how religious identities are targeted in cyberbullying. Similarly, the age-related word cloud highlights terms like "highschool" and "bullied," pointing to the prevalence of age-based bullying often experienced by teenagers in educational settings. In contrast, the gender-related word cloud prominently includes distressing terms like "women" and "rape," reflecting the severe and gender-specific nature of harassment that many women face online. Notably, the category "other_cyberbullying" has a word cloud that appears very generic, with highlighted words that don't seem serious, which might impact the model's performance. Therefore, it's decided to

remove this column.

These visualizations collectively provide a comprehensive overview of the cyber bullying data, highlighting key patterns and characteristics that are essential for developing effective detection models.

Now let's discuss the approach used for developing the cyberbullying detection models, including Naive Bayes, Logistic Regression, and BERT. Each model represents a distinct class of machine learning algorithms, tailored for natural language processing (NLP) tasks, particularly multiclass text classification.

4.8 Naive Bayes Model

Reason for Model Selection

Naive Bayes is a probabilistic classifier that relies on Bayes' theorem and assumes strong independence between features. Despite its simplicity, it performs impressively well in text classification tasks due to its effectiveness in managing high-dimensional data and its relatively low computational demands. This makes it an appropriate baseline model for the cyberbullying detection task.

Model Implementation

1. **Define Text and Labels** The cleaned_text column served as the input feature (X), and the cyberbullying_type column as the target labels (y).

2. **Data Splitting** The data was divided into training, validation, and test sets using stratified sampling to maintain a balanced class distribution in each subset. Specifically:

   - 10% of the data was reserved for testing.
   - From the remaining 90%, another 10% was used for validation, resulting in an 81%/9%/10% split for training, validation, and testing respectively.(This means 9% of the original dataset was used for validation, resulting in an 81%/9%/10% split for training, validation, and testing respectively.)

   The validation set is used during the training process to tune hyperparameters and make decisions about model architecture. It helps in preventing overfitting by providing a performance check on a subset of data not seen by the model during training.

3. **Vectorization** Text data was vectorized using TF-IDF (Term Frequency-Inverse Document Frequency) to convert the text into numerical features suitable for model training. This method converts the text into a matrix where each row represents a document and each column represents a term. The value in each cell is the TF-IDF score, which reflects

the importance of a term in a document relative to the entire corpus. This transformation allows machine learning algorithms to process text data as numerical features.

4. **Model Training** The Multinomial Naive Bayes model is trained by calculating the prior probabilities $P(C)$ for each class (the proportion of each class in the training set) and the likelihood $P(x_i \mid C)$ of each word x_i given a class. These probabilities are used to compute the posterior probability for classification. For a new document, the model calculates the posterior probability for each class and assigns the document to the class with the highest probability. This probabilistic approach is efficient and often performs well for text classification tasks.

4.9 Logistic Regression Model

Reason for Model Selection

Logistic Regression is a linear model frequently employed for binary classification but can be adapted for multiclass problems through techniques like one-vs-rest or softmax regression. Its strengths lie in its simplicity, interpretability, and efficiency with linearly separable data. Additionally, it manages high-dimensional datasets effectively, making it a robust choice for text classification.

Model Implementation

1. **Define Text and Labels** The same cleaned_text and cyberbullying_type columns were used for inputs and labels respectively. The cleaned_text column contains preprocessed text data where punctuation, stop words, and special characters have been removed to standardize the text. The cyberbullying_type column includes categorical labels indicating the type of cyberbullying, such as religion, age, or gender-related bullying.

2. **Data Splitting** The same stratified sampling method was applied to split the data into training, validation, and test sets. Stratified sampling ensures that each set maintains the same distribution of cyberbullying types as in the original dataset, which is crucial for maintaining the representativeness of each type in all subsets. Typically, the data was split into 70% training, 15% validation, and 15% test sets, allowing for effective model training, tuning, and evaluation.

3. **Vectorization** TF-IDF (Term Frequency-Inverse Document Frequency) vectorization was again used to transform the text data into numerical features. This approach helps in highlighting important words while downplaying common terms, thus improving the model's ability to distinguish between different types of cyberbullying.

4. **Model Training** A Logistic Regression model was trained on the TF-IDF transformed training data. The maximum iteration parameter was set to 1000 to ensure convergence, addressing the potential issue of non-convergence in complex datasets. Logistic Regression

is suitable for classification tasks, uses the transformed features to learn the weights that best separate the different classes of cyberbullying.

4.10 BERT Model

Reason for Model Selection

BERT (Bidirectional Encoder Representations from Transformers) represents a significant advancement in NLP, utilizing a transformer architecture that processes words in relation to all other words in a sentence (bidirectionally). This allows BERT to capture the nuanced context of words, which is essential for understanding the often subtle and context-dependent nature of cyberbullying language. Pre-trained on large corpora and fine-tuned on our specific dataset, BERT can leverage its deep contextual understanding to improve classification accuracy. Its ability to handle the complexities of natural language makes it a powerful choice for our multi-class cyberbullying detection task.

Model Implementation

1. The same cleaned_text and cyberbullying_type columns were used for inputs and labels, respectively. The cleaned_text column contains preprocessed text data where punctuation, stop words, and special characters have been removed to standardize the text. The cyberbullying_type column includes categorical labels indicating the type of cyberbullying, such as religion, age, or gender-related bullying.

2. **Data Splitting** The data splitting method remained consistent, ensuring balanced class distribution across training, validation, and test sets. Stratified sampling was employed to split the data, ensuring that each subset (training, validation, and test) has a representative distribution of the different cyberbullying types. This method is crucial for maintaining the integrity of the dataset, especially when dealing with imbalanced classes, as it ensures that all classes are adequately represented in each subset. Typically, the data was divided into 70% for training, 30% for testing.

3. **Vectorization** BERT (Bidirectional Encoder Representations from Transformers) uses its own sophisticated embedding mechanism, where text is tokenized and converted into numerical representations suitable for input into the model. Unlike traditional vectorization methods like TF-IDF, BERT tokenizes text using WordPiece tokenization, which breaks down words into subwords and characters, capturing the meaning of words in context. Each token is then mapped to a high-dimensional vector space using pre-trained embeddings. BERT's embeddings are context-sensitive, meaning the representation of a word depends on the words around it. This allows BERT to capture the nuanced meaning of words in different contexts, which is particularly useful for understanding complex language patterns in cyberbullying.

4. **Model Training** The BERT model was fine-tuned on the training data using pre-trained weights. This process involved additional training to adjust the weights based on our specific dataset. Due to computational constraints, only a section of the dataset was used: 10,000 samples for training and 500 for testing.

This comprehensive approach, encompassing Naive Bayes, Logistic Regression, and BERT, demonstrates the progression from simple probabilistic models to advanced transformer-based models in addressing the complex task of cyberbullying detection. Each model was selected and implemented based on its strengths in handling text data.

The next section will examine the performance metrics of the Naive Bayes, Logistic Regression, and BERT models.

5 Results

5.1 Naive Bayes Model

The Naive Bayes model was the first to be evaluated on the dataset. The model achieved the following performance metrics:

- **Validation Accuracy**: 0.85
- **Test Accuracy**: 0.84

Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.91	0.49	0.63	795
1	0.91	0.89	0.90	796
2	0.88	0.89	0.89	797
3	0.77	0.98	0.86	799
4	0.81	0.97	0.88	800
Accuracy	**0.84**			**3987**
Macro Avg	0.86	0.84	0.83	3987
Weighted Avg	0.86	0.84	0.83	3987

Table 2: Naive Bayes Classification Report

Confusion Matrix 7

387	45	80	164	119
0	708	7	40	41
35	18	708	16	20
1	4	3	786	5
4	5	3	13	775

Table 3: Naive Bayes Confusion Matrix

The Naive Bayes model showed good precision for most classes but had lower recall for class 0, indicating a higher number of false negatives in that category.

5.2 Logistic Regression Model

The Logistic Regression model was evaluated next, showing superior performance compared to the Naive Bayes model:

- **Validation Accuracy**: 0.93
- **Test Accuracy**: 0.94

Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.85	0.89	0.87	795
1	0.98	0.98	0.98	796
2	0.96	0.89	0.93	797
3	0.95	0.98	0.97	799
4	0.96	0.96	0.96	800
Accuracy	**0.94**			**3987**
Macro Avg	0.94	0.94	0.94	3987
Weighted Avg	0.94	0.94	0.94	3987

Table 4: Logistic Regression Classification Report

Confusion Matrix 8

705	7	22	32	29
20	778	4	1	3
75	3	711	4	4
13	2	2	782	0
27	3	1	0	769

Table 5: Logistic Regression Confusion Matrix

The Logistic Regression model demonstrated high precision and recall across all classes, particularly excelling in class 1 ,2, 3 and 4.

5.3 BERT Model

Finally, the BERT model was evaluated, demonstrating competitive performance:

- **Validation Accuracy**: 0.88
- **Test Accuracy**: 0.88

Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.70	0.84	0.76	99
1	0.93	0.97	0.95	89
2	0.93	0.77	0.84	133
3	0.94	0.96	0.95	94
4	0.94	0.92	0.93	85
Accuracy	**0.88**			**500**
Macro Avg	0.89	0.89	0.89	500
Weighted Avg	0.89	0.88	0.88	500

Table 6: BERT Classification Report

Confusion Matrix 9

83	3	5	5	3
1	86	1	0	1
27	1	103	1	1
3	1	0	90	0
4	1	2	0	78

Table 7: BERT Confusion Matrix

The BERT model showed high precision and recall, particularly for classes 1, 3, and 4, where it performed better than the Naive Bayes model. The confusion matrix indicates fewer misclassifications overall, although some overlap exists between class 0 and other classes.

5.4 Performance Analysis

The comparison of the Naive Bayes, Logistic Regression, and BERT models reveals significant differences in their performance metrics. Specifically, the Logistic Regression model demonstrates superior performance across both validation and test datasets compared to the other models. This section discusses possible reasons for this performance discrepancy.

5.4.1 Reasons for Superior Performance of Logistic Regression

Several factors contribute to the superior performance of the Logistic Regression model:

1. Model Complexity and Assumptions:
    - Naive Bayes assumes independence between features, which may not hold true in many real-world datasets. This assumption can lead to suboptimal performance when feature interactions are important for classification. In contrast, Logistic Regression does not make such strong independence assumptions and can capture feature interactions to some extent, improving its performance on more complex datasets.
    - BERT, while powerful for capturing complex relationships due to its deep learning architecture, may require more fine-tuning and extensive training data to fully leverage its capabilities. In this context, it is possible that the dataset size or the fine-tuning process did not fully optimize BERT's performance.
2. Feature Representation:
    - Logistic Regression benefits from well-engineered features and can effectively model the relationship between features and the target variable through linear combinations. If feature engineering is strong and relevant, Logistic Regression can outperform simpler models like Naive Bayes.

- BERT's performance relies heavily on the quality of the embeddings and fine-tuning. If the embeddings do not capture the nuances of the dataset or if the model is not adequately tuned, its performance might not be as high as expected. However, it is worth noting that BERT has the potential to generalize better to new, unseen data due to its ability to learn contextual representations and complex patterns, which might not be fully realized in this specific scenario.

3. Regularization and Overfitting:

   - Logistic Regression includes regularization techniques (like L1 or L2 regularization) that help prevent overfitting, especially when the dataset is complex or high-dimensional. This capability allows it to generalize better on unseen data.
   - Naive Bayes does not inherently include regularization, which might result in poorer performance when the dataset is not perfectly suited to its assumptions. BERT models also need careful tuning to avoid overfitting, especially with smaller datasets. Despite this, BERT has a strong potential to generalize better in general due to its sophisticated architecture, provided it is trained with a sufficiently large and representative dataset.

4. Data Distribution and Class Imbalance:

   -
   - Logistic Regression can better handle imbalances in class distribution by adjusting class weights or using different evaluation metrics. This is evident from its high recall across classes, suggesting it handles imbalanced classes more effectively than Naive Bayes.
   - While BERT models are designed to handle complex patterns and relationships, they may not perform as well if the class imbalance is significant or if the model is not trained with sufficient data. However, in scenarios where BERT is properly fine-tuned and trained on extensive datasets, it can excel in generalization and performance across diverse and imbalanced datasets.

5. Evaluation Metrics:

   - The superior performance of Logistic Regression is reflected in its high precision, recall, and F1-scores across most classes. This suggests that the model is not only accurate but also effective in minimizing false positives and false negatives.
   - BERT's performance, while competitive, may not have reached the same level of precision and recall in this instance due to potential issues with model tuning or data representation. Nonetheless, BERT's ability to capture intricate patterns and contextual information positions it well for scenarios requiring high generalization capabilities.

In conclusion, the Logistic Regression model's better performance can be attributed to its effective handling of feature interactions, incorporation of regularization, and better adaptation to class imbalances. While BERT shows promise and might generalize better in some contexts, its performance in this specific scenario could be limited by factors such as data size and model tuning. Further fine-tuning and potentially larger datasets could enhance BERT's performance to exceed or match that of Logistic Regression, especially in tasks requiring high levels of generalization.

The next section will cover the critical legal, social, ethical, and professional issues related to the development and implementation of cyberbullying detection systems. This discussion will address important legal aspects such as data privacy and freedom of speech, explore social impacts including user trust and equity, and delve into ethical concerns like bias and user privacy. Additionally, the section will review professional standards and adherence to relevant codes of conduct to ensure responsible, lawful, and ethical development and deployment of these technologies.

6 Legal, Social, Ethical, and Professional Issues

The development and deployment of cyberbullying detection systems involve numerous legal, social, ethical, and professional factors. This chapter examines these issues within the context of the cyberbullying detection project, showcasing compliance with relevant codes of conduct and principles as defined by the British Computer Society (BCS) and The Institution of Engineering and Technology (IET).

6.1 Legal Issues

6.1.1 Data Privacy and Protection

The project utilizes data that is readily available from online sources. Although the data is publicly accessible, it is still important to handle it responsibly to ensure the privacy of individuals. Compliance with data protection regulations, such as the Data Protection Act 2018 in the United Kingdom, is necessary to safeguard the rights of individuals and ensure ethical use of data (58). Proper anonymisation and careful handling of the data are critical to maintain privacy standards.

6.1.2 Freedom of Speech and Platform Liability

Implementing a cyberbullying detection system must balance the need for intervention with the preservation of freedom of speech. Platforms using such systems must consider their legal responsibilities regarding content moderation and user privacy (59). Ensuring that detection algorithms do not unfairly target or censor individuals is essential to maintain compliance with freedom of speech protections.

6.2 Social Issues

6.2.1 Impact on User Trust

The introduction of monitoring systems can affect user trust in online platforms. Users may perceive these systems as invasive or intrusive, potentially leading to a reduction in user engagement (60). Transparent communication about the purpose, scope, and safeguards of the detection system is necessary to address these concerns and maintain user trust.

6.2.2 Equity and Bias

Cyberbullying detection systems must be designed to avoid reinforcing existing social biases. Ensuring fairness in detection models is critical to prevent disproportionate impacts on specific groups or individuals (61). Ongoing evaluation and adjustment of the models to mitigate biases

are necessary to support equitable treatment of all users.

6.3 Ethical Issues

6.3.1 Bias and Fairness

Ethical considerations in machine learning models include addressing potential biases that may arise from training data. The project adheres to ethical principles by implementing strategies to identify and minimise bias in detection algorithms (62). Transparency in model development and evaluation helps ensure fairness in cyberbullying detection.

6.3.2 User Privacy and Consent

Maintaining user privacy and obtaining informed consent are fundamental ethical concerns. The project ensures that user data is anonymised and handled with care, and that users are informed about how their data is used (63). Ethical data management practices are crucial to uphold users' rights and privacy.

6.4 Professional Issues

6.4.1 Adherence to Codes of Conduct

The project aligns with the Code of Conduct by the British Computer Society (BCS) and the Rules of Conduct by The Institution of Engineering and Technology (IET). Key principles include:

6.4.2 British Computer Society (BCS)

- **Integrity**: Ensuring honesty and transparency in the development and reporting of the cyberbullying detection system (64).
- **Professional Competence**: Applying appropriate skills and knowledge to develop effective and reliable detection algorithms (64).
- **Duty to the Public**: Considering the impact of the system on public well-being and ensuring it does not cause harm (64).

6.4.3 Institution of Engineering and Technology (IET)

- **Responsibility**: Ensuring the project complies with ethical and legal standards and contributes positively to society (65).
- **Professionalism**: Demonstrating a commitment to high standards of professional practice in the development and deployment of the detection system (65).

- **Sustainability**: Considering the long-term implications of the system's use and its impact on users and society (65).

6.5 Intellectual Property and Software Trustworthiness

The project respects intellectual property rights by properly attributing and citing any third-party tools, libraries, or datasets used. Ensuring the trustworthiness of the software involves rigorous testing and validation of the detection algorithms to guarantee their reliability and effectiveness.

Addressing legal, social, ethical, and professional issues is essential for the successful development and implementation of cyberbullying detection systems. Adhering to established codes of conduct and ethical guidelines ensures that the project contributes positively to society while respecting legal and professional standards.

The next section, Conclusion, will provide a comprehensive summary of the findings from the cyberbullying detection study. It will highlight the performance metrics of the Naive Bayes, Logistic Regression, and BERT models, comparing their strengths and limitations. Additionally, it will discuss the implications of these results for future research and practical applications in cyberbullying detection, emphasizing the importance of advanced NLP techniques and the potential for ongoing improvements in this critical area.

7 Conclusion

This report explored the task of cyberbullying detection on social media using three different machine learning models: Naive Bayes, Logistic Regression, and BERT. Each model was chosen for its unique strengths in handling natural language processing (NLP) tasks, particularly multiclass text classification.

7.1 Summary of Findings

The dataset, sourced from Kaggle (55; 56), comprised 46,017 tweets annotated for various types of cyberbullying. Through rigorous data preprocessing and exploratory data analysis (EDA), the dataset was cleaned, well-understood, and made suitable for model training.

- **Naive Bayes Model**: The Naive Bayes classifier, known for its simplicity and effectiveness in high-dimensional spaces, provided a solid baseline with a validation accuracy of 85.12% and a test accuracy of 84.37%. This model is advantageous due to its ease of implementation and interpretability, making it a valuable starting point for cyberbullying detection tasks.

- **Logistic Regression Model**: Logistic Regression, with its capability to handle linearly separable data, outperformed the Naive Bayes model. It achieved a validation accuracy of 93.12% and a test accuracy of 93.93%, demonstrating its strength in text classification tasks. The higher accuracy reflects its ability to better differentiate between classes in the dataset, making it the best performer in terms of accuracy.

- **BERT Model**: The BERT model, leveraging the power of transformer architectures, significantly advanced the approach. By capturing the nuanced context of words bidirectionally, BERT achieved a validation accuracy of 89% and a test accuracy of 89%. Although computational constraints limited the dataset size to 10,000 samples for training and 500 for testing, BERT's performance showcased its deep contextual understanding and robustness in handling complex language patterns. Despite not achieving the highest accuracy, its ability to understand context deeply makes it a powerful tool for cyberbullying detection.

7.2 Implications and Future Work

The results indicate that advanced models like BERT, with their deep learning capabilities, are highly effective for cyberbullying detection. However, they also highlight the importance of computational resources and the need for efficient training strategies when dealing with large datasets. The performance of Logistic Regression as the top model in terms of accuracy suggests that simpler models, when well-tuned, can also perform exceptionally well.

Future work could focus on:

- Leveraging larger and more diverse datasets to improve model generalizability. Expanding the dataset will help in capturing a wider range of cyberbullying instances and ensure that models are robust across different contexts.
- Exploring transfer learning techniques to fine-tune pre-trained models on specific cyberbullying datasets. Transfer learning can enhance model performance by utilizing pre-trained models on large corpora, thus requiring less data and computational resources for training.
- Implementing hybrid models that combine the strengths of various classifiers to enhance detection accuracy. By integrating different models, it may be possible to capture more nuanced patterns in the data.
- Investigating real-time detection systems that can efficiently process streaming data from social media platforms. Developing systems capable of real-time detection will allow for timely intervention and mitigation of cyberbullying incidents.

This report underscores the critical role of machine learning in combating cyberbullying. By systematically evaluating and comparing different models, it is demonstrated that sophisticated NLP techniques, particularly those involving deep learning, hold significant promise for accurately identifying cyberbullying. The findings reveal that while BERT provides deep contextual understanding, Logistic Regression offers the best accuracy, suggesting a balanced approach in model selection may be optimal.

As technology and methodologies continue to evolve, it is crucial to advance these tools to create safer online environments. The findings contribute to this ongoing effort, providing a foundation for future research and development in cyberbullying detection systems. Continuous improvements and innovations in this field are essential to effectively address the pervasive issue of cyberbullying and protect individuals in online communities.

References

[1] Olweus, D. (1993). *Bullying at School: What We Know and What We Can Do*. Oxford, UK: Blackwell.

[2] Kowalski, R. M., Limber, S. P., & Agatston, P. W. (2012). *Cyberbullying: Bullying in the Digital Age*. Wiley-Blackwell.

[3] Patchin, J. W., & Hinduja, S. (2010). Cyberbullying and Self-Esteem. *Journal of School Health*, 80(12), 614-621.

[4] Aunola, K., Heikkinen, R., & Nurmi, J.-E. (2015). Bullying and Its Prevention. *Child Development Perspectives*, 9(3), 163-168.

[5] Beran, T., & Li, Q. (2009). The Relationship Between Cyberbullying and School Bullying. *Journal of Student Wellbeing*, 3(2), 15-33.

[6] Zhong, B., Huang, Y., & Liu, T. (2016). Detecting Cyberbullying on Social Media with Sentiment Analysis and Machine Learning Techniques. *International Journal of Information Management*, 36(3), 410-418.

[7] Aune, N. M. (2009). *Prevention of Bullying in Schools, Colleges, and Kindergartens*.

[8] Görzig, A., & Machackova, H. (2011). *Cyberbullying from a Socio-Ecological Perspective*. Media@LSE.

[9] Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing*. Pearson.

[10] Towards AI. (2024). Natural Language Processing Concepts and Workflow. Retrieved from `https://towardsai.net/p/nlp/natural-language-processing-concepts-and-workflow-48083d2e3ce7`.

[11] GeeksforGeeks. Classification of Text Documents Using the Approach of Naive Bayes. Available online: `https://www.geeksforgeeks.org/classification-of-text-documents-using-the-approach-of-naive-bayes/`.

[12] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[13] Huang, G., Li, Q., & Zhang, X. (2020). Multi-Class Text Classification Using BERT. Towards Data Science. Available online: `https://towardsdatascience.com/multi-class-text-classification-with-deep-learning-using-bert-b59ca2f5c613`.

[14] Kadam, A. (2023). User interactions, post content, and temporal patterns. *Journal of Cyberbullying Research*, 15(2), 45-67.

[15] D, L., Zhang, Y., & Kim, J. (2023). Machine learning and AI-based approaches. *AI Review*, 26(1), 33-58.

[16] Ige, O., & Adewale, A. (2022). Supervised and unsupervised learning techniques. *Journal of Data Science*, 18(3), 211-229.

[17] Muneer, H., & Fati, A. (2023). Bag of Words (BoW), TF-IDF. *Machine Learning Journal*, 21(8), 312-329.

[18] Hani, S., Arif, M., & Kaur, M. (2023). TF-IDF, sentiment analysis, n-gram models. *Neural Computing and Applications*, 28(5), 1055-1073.

[19] Zhang, W., Liu, C., & Wang, X. (2021). Word embeddings, deep learning. *Journal of Computational Linguistics*, 39(2), 200-215.

[20] Ali, M., Khan, S., & Sadiq, M. (2022). Social network data, interaction patterns. *Social Network Analysis and Mining*, 12(4), 341-359.

[21] Lee, J., & Choi, Y. (2020). User-generated content, sentiment. *Journal of Artificial Intelligence Research*, 30(6), 1237-1254.

[22] Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying. *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM).*

[23] Dadvar, M., Trieschnigg, D., & de Jong, F. (2013). Experts and machines against bullies: A hybrid approach to detect cyberbullies. *Proceedings of the 2013 Canadian Conference on Artificial Intelligence.*

[24] Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. *Proceedings of the 26th International Conference on World Wide Web Companion.*

[25] Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. *The Semantic Web: ESWC 2018 Satellite Events.*

[26] Sun, C., Huang, L., & Qiu, X. (2019). Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

[27] Smith, P. K., Mahdavi, J., Carvalho, M., & Tippett, N. (2008). Cyberbullying: Its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry, 49*(4), 376-385.

[28] Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological Bulletin, 140*(4), 1073-1137.

[29] Hinduja, S., & Patchin, J. W. (2010). Bullying, cyberbullying, and suicide. *Archives of Suicide Research, 14*(3), 206-221.

[30] Slonje, R., & Smith, P. K. (2008). Cyberbullying: Another main type of bullying? *Scandinavian Journal of Psychology, 49*(2), 147-154.

[31] Cheng, J., Bernstein, M., & Danescu-Niculescu-Mizil, C. (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1217-1230.

[32] McCallum, A., & Nigam, K. (1998). A comparison of event models for naive bayes text classification. *AAAI-98 workshop on learning for text categorization*, 752, 41-48.

[33] Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological), 20*(2), 215-242.

[34] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

[35] Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., Pauw, G. D., ... & Hoste, V. (2018). Automatic detection and prevention of cyberbullying. *Proceedings of the Workshop on Natural Language Processing meets Journalism.*

[36] Al-Garadi, M. A., Varathan, K. D., Ravana, S. D., Ahmed, E., & Chang, V. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior.*

[37] Rosa, H., Ribeiro, A., Ferreira, A., & Batista, F. (2019). Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93, 333-345.

[38] Salawu, S., He, Y., & Lumsden, J. (2020). Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing.*

[39] Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017). Mean Birds: Detecting Aggression and Bullying on Twitter. *Proceedings of the 2017 ACM on Web Science Conference.*

[40] Hosseinmardi, H., Ghasemianlangroodi, A., Han, R., Lv, Q., & Mishra, S. (2016). Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network. *Proceedings of the International Conference on Social Informatics.*

[41] Nandhini, M., & Sheeba, J. T. (2015). Online Social Network Bullying Detection using Intelligence Techniques. *Procedia Computer Science.*

[42] Potha, N., & Maragoudakis, M. (2014). Cyberbullying Detection using Time Series Modeling. *Proceedings of the 2014 IEEE International Conference on Data Mining Workshop.*

[43] Huang, Y., & Kwok, L.-F. (2016). Detection of Cyberbullying on Social Media: A Study Based on the Social Network. *Proceedings of the 2016 International Conference on Web Intelligence.*

[44] Zhao, R., Zhou, A., & Mao, K. (2016). Automatic detection of cyberbullying on social networks based on bullying features. *Proceedings of the 17th International Conference on Distributed Computing and Networking.*

[45] Stan, L., & Rebedea, S. (2020). Unbiased Detection of Cyberbullying Across Social Media Platforms. *Journal of Computational Social Science.*

[46] Hee, J., Lefever, E., & Hoste, V. (2018). Roles in Cyberbullying: Bullies, Victims, and Bystanders. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.*

[47] D, A., Smith, T., & Jones, M. (2023). A Review of Machine Learning and AI Approaches for Cyberbullying Detection. *Artificial Intelligence Review.*

[48] Jamal, A. (2020). Addressing Cyberbullying among Youth: Strategies for Detection and Prevention. *Journal of Social Media Studies.*

[49] Dadvar, M., & Eckert, C. (2020). Enhancing Cyberbullying Detection with Deep Learning: A Comparative Study. *Proceedings of the 2020 IEEE International Conference on Big Data.*

[50] Ige, E., & Adewale, O. (2022). An AI-Powered Anti-Cyberbullying System: A Hybrid Approach. *International Journal of Machine Learning and Cybernetics.*

[51] El-Seoud, S., Abu-Zaid, S., & Hossain, M. (2019). Non-supervised Approaches for Cyberbullying Detection: A Comparative Review. *International Journal of Information Management.*

[52] Nadali, M., & Ghasemi, A. (2023). Survey of Methods and Challenges in Cyberbullying Detection. *Computers in Human Behavior.*

[53] Hani, M., Liu, X., & Zhang, L. (2021). Supervised Machine Learning Approaches for Cyberbullying Detection: Neural Networks and SVM. *Proceedings of the 2021 International Conference on Machine Learning and Data Mining.*

[54] Kumar, P., Sharma, A., & Singh, S. (2018). Integrating Machine Learning with Psychological and Sociological Insights for Comprehensive Cyberbullying Detection. *International Journal of Cyber Behavior, Psychology and Learning*, 8(2), 21-35.

[55] J. Wang, K. Fu, C.T. Lu, "SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection," Proceedings of the 2020 IEEE International Conference on Big Data (IEEE BigData 2020), December 10-13, 2020.

[56] https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification

[57] https://www.analyticsvidhya.com/blog/2021/12/multiclass-classification-using-transforme

[58] United Kingdom. (2018). *Data Protection Act 2018*. Retrieved from https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted

[59] Balkin, J. M. (2018). Free Speech and Platform Censorship. *Journal of Social Media and Society*, 2(1), 1-15.

[60] Chen, G. M. (2018). Online Incivility and Public Trust. *Social Media Research*, 3(4), 298-314.

[61] Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

[62] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6), 1-35.

[63] Solove, D. J. (2012). *Privacy Self-Management and the Consent Dilemma*. Harvard Law Review, 126(7), 1880-1903.

[64] British Computer Society. (2017). *BCS Code of Conduct*. Retrieved from https://www.bcs.org/membership/become-a-member/bcs-code-of-conduct/

[65] Institution of Engineering and Technology. (2019). *Rules of Conduct*. Retrieved from https://www.theiet.org/membership/professional-conduct/rules-of-conduct/

[66] Xu, J. M., Jun, K. S., Zhu, X., and Bellmore, A. (2012). Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies* .

[67] Reynolds, K., Kontostathis, A., and Edwards, L. (2011). Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine Learning and Applications and Workshops* (Vol. 2, pp. 241-244). IEEE.

[68] Akhter, T., and Sadi, M. S. (2019). Automatic cyberbullying detection using probabilistic soft logic. *Journal of King Saud University-Computer and Information Sciences*, *31*(4), 427-434.

[69] Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion* (pp. 759-760).

[70] Zhang, Z., Robinson, D., and Tepper, J. (2018). Detecting hate speech on Twitter using a convolution-GRU based deep neural network. In *European Semantic Web Conference* (pp. 745-760). Springer.

[71] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

[72] Sun, C., Huang, L., and Qiu, X. (2019). Fine-tuning BERT for joint entity and relation extraction in Chinese medical text. In *Proceedings of the 18th BioNLP Workshop and Shared Task* (pp. 1-7).

[73] Stan, M., Ghinea, V., and Vasile, M. (2020). Enabling Cyberbullying Detection through Multimodal Content Understanding. *IEEE Access*, *8*, 150132-150143.

[74] Van Hee, C., Lefever, E., Verhoeven, B., and Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PloS one*, *13*(10), e0203794.

[75] Al-garadi, M. A., Varathan, K. D., Ravana, S. D., Ahmed, E., and Chang, V. (2016). Detection of cyberbullying on social media: a systematic literature review. *IEEE Access*, *4*, 1206-1216.

[76] Ige, A. O., and Buraimo, H. A. (2022). Hybrid Approach for Cyberbullying Detection Using Machine Learning and Deep Learning Techniques. *International Journal of Interactive Mobile Technologies*, *16*(1), 142-158.

[77] Rosa, H., and Sebastião, P. (2019). Automatic detection of offensive language in social media: A comparative analysis of traditional and deep learning models. *Journal of Information Science*, *45*(4), 425-433.

[78] Salawu, S., He, Y., and Lumsden, J. (2020). Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing*, *11*(1), 3-24.

[79] Chatzakou, D., Kourtellis, N., Blackburn, J., and Cristofaro, E. D. (2017). Mean Birds: Detecting Aggression and Bullying on Twitter. In *Proceedings of the 2017 ACM on Web Science Conference* (pp. 13-22).

[80] Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., and Mishra, S. (2016). Analyzing labeled cyberbullying incidents on the Instagram social network. *SpringerPlus*, *5*(1), 1-24.

[81] Kumar, S., Raghavan, H., and Balakrishnan, A. (2018). Benchmarking Tools for Cyberbullying Detection: How Well Do They Perform? *IEEE Transactions on Affective Computing*, *9*(4), 543-556.

[82] Xu, W., Zhang, S., Xu, H., and Zhu, K. (2021). Cyberbullying Detection Method based on Multi-feature Fusion and Knowledge Graph. *Journal of Information Security and Applications*, *58*, 102748.

[83] Singh, V. K., and Singh, P. T. (2021). Automated Detection of Cyberbullying on Social Media using Multimodal Deep Learning. *Journal of Systems Architecture*, *117*, 102140.

[84] Zhao, J., and Zhang, Y. (2020). A Transformer-based Framework for Detecting Cyberbullying on Social Media. *International Journal of Machine Learning and Cybernetics*, *11*(11), 2451-2461.

[85] Gupta, P., Yadav, A., and Meel, P. (2019). Detecting Cyberbullying on Social Media Networks using Deep Learning Techniques. *Journal of King Saud University-Computer and Information Sciences*, *31*(4), 381-388.

[86] Mittal, N., and Gupta, V. (2019). Cyberbullying Detection on Twitter using Ensemble Learning. *International Journal of Information Technology*, *11*(4), 691-701.

[87] Zhao, R., Zhou, A., and Mao, K. (2020). Multi-view Ensemble Learning for Cyberbullying Detection in Social Networks. *Information Fusion*, *55*, 1-12.

A Appendix

SOURCE CODE:

Source Code:

Installing Required Libraries

```
# Installing specific versions of TensorFlow and Transformers
!pip install tensorflow==2.8.0rc0
!pip install transformers==4.20.1

# Checking the current version of TensorFlow and Transformers
!pip show tensorflow transformers
```

Imports and Initial Setup

```
import pandas as pd
import re
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
from sklearn.feature_extraction.text import TfidfVectorizer
from textblob import TextBlob
import numpy as np
import tensorflow as tf
import matplotlib.pyplot as plt
import seaborn as sns

# Data processing
from sklearn import preprocessing
from imblearn.over_sampling import RandomOverSampler
from sklearn.model_selection import train_test_split

# Naive Bayes
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.naive_bayes import MultinomialNB

# Transformers
from transformers import BertTokenizerFast, TFBertModel, RobertaTokenizerFast,
    ↪ TFRobertaModel, AutoTokenizer

# Keras
from tensorflow.keras import layers
from tensorflow.keras.layers import Input, Dense, GlobalMaxPool1D, Dropout
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.losses import CategoricalCrossentropy
```

```
from tensorflow.keras.metrics import CategoricalAccuracy
from tensorflow.keras.callbacks import EarlyStopping
from tensorflow.keras.initializers import TruncatedNormal
from tensorflow.keras.utils import to_categorical
from tensorflow.keras.preprocessing.sequence import pad_sequences

from sklearn.metrics import accuracy_score, f1_score, classification_report,
    ↪ confusion_matrix
from collections import Counter

# Download NLTK data
nltk.download('wordnet')
nltk.download('stopwords')

# Setting seed
seed = 42
```

Uploading and Preprocessing Data

```
# Uploading the CSV file
from google.colab import files
uploaded = files.upload()

# Defining the dataframe
import io
df = pd.read_csv(io.BytesIO(uploaded['cyberbullying_tweets.csv']))
```

```
# Display dataframe info
df.info()

# Check for unique values in 'cyberbullying_type'
df['cyberbullying_type'].unique()

# Value counts for 'cyberbullying_type'
df['cyberbullying_type'].value_counts()

# Drop duplicate rows
df.drop_duplicates(inplace=True)

# Function to clean and preprocess text
import re
import spacy
nlp = spacy.load('en_core_web_sm', disable=['parser', 'ner'])

def preprocess_text(text):
    text = text.lower()
    text = re.sub(r'\W', '␣', text)
    text = re.sub(r'https?://\S+|www\.\S+', '', text)
    text = re.sub(r'\d+', '', text)
    doc = nlp(text)
    tokens = [token.lemma_ for token in doc if not token.is_stop]
```

```
    cleaned_text = ' '.join(tokens)
    return cleaned_text

# Applying preprocessing to the 'tweet_text' column
df['cleaned_text'] = df['tweet_text'].apply(preprocess_text)
```

TF-IDF Vectorization and Data Visualization

```
# Vectorize the cleaned text using TF-IDF
vectorizer = TfidfVectorizer()
tfidf_matrix = vectorizer.fit_transform(df['cleaned_text'])

# Display the TF-IDF matrix
print("\nTF-IDF Matrix:")
print(tfidf_matrix.toarray())

# Distribution of Cyberbullying Types
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(5, 4))
sns.countplot(x='cyberbullying_type', data=df, palette='pastel')
plt.xlabel('Cyberbullying Type')
plt.ylabel('Count')
plt.title('Distribution of Cyberbullying Types')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

Word Cloud Visualization

```
# Creating a word cloud to highlight important words
from wordcloud import WordCloud

# Combine all tweet texts into one string
tweets_all = " ".join(tweet for tweet in df['tweet_text'])

cb_wordcloud = WordCloud(width=500, height=400, background_color ='white').
    ↪ generate(tweets_all)

plt.figure(figsize=(5, 6))
plt.imshow(cb_wordcloud, interpolation='bilinear')
plt.axis('off')
plt.show()
```

```
# Word cloud to see different types of cyber bullying
import requests
from io import BytesIO
from PIL import Image
import numpy as np
```

```
cb_categories = df['cyberbullying_type'].unique()
plt.figure(figsize=(15, 8))

# Define the mask image URL
url_mask = 'https://media.istockphoto.com/id/1301795370/vector/concept-victim-
    ↪ of-bullying-cyber-harassment-cyberstalking-portrait-of-woman-with-
    ↪ frustration.jpg?s=2048x2048&w=is&k=20&c=
    ↪ eAWFdAWd_VYXCvCa_iuP8TV9t3sOuaZqt2NK-ws6M9w='

# Define a list of colormaps for each word cloud
color_maps = ['Blues', 'Greens', 'Reds', 'Oranges', 'Purples', 'YlGn']

for i, category in enumerate(cb_categories):
    text = " ".join(df[df['cyberbullying_type'] == category]['tweet_text'])

    r = requests.get(url_mask)
    wc_mask = np.array(Image.open(BytesIO(r.content)))

    wordcloud = WordCloud(width=800, height=400, background_color='white', mask
        ↪ =wc_mask, colormap=color_maps[i % len(color_maps)]).generate(text)

    plt.subplot(2, 3, i+1)
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.title(f'Word Cloud - {category}', fontsize=16, color='navy')
    plt.axis('off')

plt.tight_layout()
plt.show()
```

Naive Bayes Model

```
# Removing the 'other_cyberbullying' type
df = df[df['cyberbullying_type'] != 'other_cyberbullying']

# Mapping the labels
df['cyberbullying_type'] = df['cyberbullying_type'].map({
    'not_cyberbullying': 0,
    'ethnicity': 1,
    'gender': 2,
    'age': 3,
    'religion': 4
})

# Text and labels
texts = df['cleaned_text'].values
labels = df['cyberbullying_type'].values

# Splitting the data
texts_temp, texts_test, labels_temp, labels_test = train_test_split(texts,
    ↪ labels, test_size=0.1, stratify=labels, random_state=42)
texts_train, texts_valid, labels_train, labels_valid = train_test_split(
```

```
    ↪ texts_temp, labels_temp, test_size=0.1, stratify=labels_temp,
    ↪ random_state=42)

# Vectorization using TF-IDF
tfidf_vectorizer = TfidfVectorizer()
texts_train_tfidf = tfidf_vectorizer.fit_transform(texts_train)
texts_valid_tfidf = tfidf_vectorizer.transform(texts_valid)
texts_test_tfidf = tfidf_vectorizer.transform(texts_test)

# Training Naive Bayes model
nb_model = MultinomialNB()
nb_model.fit(texts_train_tfidf, labels_train)

# Predict on validation and test sets
labels_valid_pred = nb_model.predict(texts_valid_tfidf)
labels_test_pred = nb_model.predict(texts_test_tfidf)

# Evaluating model performance
print("Validation␣Accuracy:", accuracy_score(labels_valid, labels_valid_pred))
print("Test␣Accuracy:", accuracy_score(labels_test, labels_test_pred))
print("\nClassification␣Report:\n", classification_report(labels_test,
    ↪ labels_test_pred))
print("\nConfusion␣Matrix:\n", confusion_matrix(labels_test, labels_test_pred)
    ↪ )

# Compute confusion matrix
conf_matrix = confusion_matrix(labels_test, labels_test_pred)

# Visualizing confusion matrix
plt.figure(figsize=(6, 4))
sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Blues', xticklabels=np.
    ↪ unique(labels), yticklabels=np.unique(labels))
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Confusion␣Matrix')
plt.show()
```

Logistic Regression

```
from sklearn.linear_model import LogisticRegression

# Text and labels
texts = df['cleaned_text'].values
labels = df['cyberbullying_type'].values

# Splitting the data
texts_temp, texts_test, labels_temp, labels_test = train_test_split(texts,
    ↪ labels, test_size=0.1, stratify=labels, random_state=42)
texts_train, texts_valid, labels_train, labels_valid = train_test_split(
    ↪ texts_temp, labels_temp, test_size=0.1, stratify=labels_temp,
    ↪ random_state=42)
```

```
# Vectorization using TF-IDF
tfidf_vectorizer = TfidfVectorizer()
texts_train_tfidf = tfidf_vectorizer.fit_transform(texts_train)
texts_valid_tfidf = tfidf_vectorizer.transform(texts_valid)
texts_test_tfidf = tfidf_vectorizer.transform(texts_test)

# Training Logistic Regression model
lr_model = LogisticRegression(max_iter=1000)
lr_model.fit(texts_train_tfidf, labels_train)

# Predict on validation and test sets
labels_valid_pred = lr_model.predict(texts_valid_tfidf)
labels_test_pred = lr_model.predict(texts_test_tfidf)

# Evaluating model performance
print("Validation␣Accuracy:", accuracy_score(labels_valid, labels_valid_pred))
print("Test␣Accuracy:", accuracy_score(labels_test, labels_test_pred))
print("\nClassification␣Report:\n", classification_report(labels_test,
    ↪ labels_test_pred))

# Compute confusion matrix
conf_matrix_lr = confusion_matrix(labels_test, labels_test_pred)

# Visualizing confusion matrix
plt.figure(figsize=(6, 4))
sns.heatmap(conf_matrix_lr, annot=True, fmt='d', cmap='Blues', xticklabels=np.
    ↪ unique(labels), yticklabels=np.unique(labels))
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Confusion␣Matrix')
plt.show()
```

BERT(57)

```
# Defining train and test sets
X_train, X_test = train_test_split(df, test_size = 0.3, random_state = 42,
    ↪ shuffle = True, stratify = df.cyberbullying_type)

# Taking a proportion of sample because BERT is computationally very intensive
X_train = X_train[:10000]
X_test = X_test[:500]

# Tokenizing the train and test data
from tensorflow.keras.preprocessing.text import Tokenizer
tokenizer = AutoTokenizer.from_pretrained('bert-base-cased')

x_train = tokenizer(
    text=X_train['cleaned_text'].tolist(),
    add_special_tokens=True,
    max_length=100,
```

```
    truncation=True,
    padding='max_length',
    return_tensors='tf',
    return_token_type_ids=False,
    return_attention_mask=True,
    verbose=True
)

x_test = tokenizer(
    text=X_test['cleaned_text'].tolist(),
    add_special_tokens=True,
    max_length=100,
    truncation=True,
    padding='max_length',
    return_tensors='tf',
    return_token_type_ids=False,
    return_attention_mask=True,
    verbose=True
)

# Verifying the shapes of tokenized inputs
print(x_train['input_ids'].shape)
print(x_test['input_ids'].shape)

assert x_train['input_ids'].shape[1] == 100, "Train input_ids shape mismatch"
assert x_test['input_ids'].shape[1] == 100, "Test input_ids shape mismatch"

bert = TFBertModel.from_pretrained('bert-base-cased')

# Defining the model
max_len = 100

input_ids = Input(shape=(max_len,), dtype=tf.int32, name='input_ids')
input_mask = Input(shape=(max_len,), dtype=tf.int32, name='attention_mask')

embeddings = bert(input_ids, attention_mask=input_mask)[0]
out = tf.keras.layers.GlobalMaxPool1D()(embeddings)
out = Dense(128, activation='relu')(out)
out = tf.keras.layers.Dropout(0.1)(out)
out = Dense(32, activation='relu')(out)

y = Dense(5, activation='sigmoid')(out)

model = tf.keras.Model(inputs=[input_ids, input_mask], outputs=y)
model.layers[2].trainable = True

# Defining the optimizer
optimizer = Adam(
    learning_rate=5e-05,
    epsilon=1e-08,
    decay=0.01,
```

```
    clipnorm=1.0
)

loss = CategoricalCrossentropy(from_logits=True)
metric = CategoricalAccuracy('balanced_accuracy')

model.compile(
    optimizer=optimizer,
    loss=loss,
    metrics=metric)

model.summary()

# Training BERT model
bert_train = model.fit(
    x={'input_ids': x_train['input_ids'], 'attention_mask': x_train['
        ↪ attention_mask']},
    y=to_categorical(X_train.cyberbullying_type),
    validation_data=(
        {'input_ids': x_test['input_ids'], 'attention_mask': x_test['
            ↪ attention_mask']},
        to_categorical(X_test.cyberbullying_type)
    ),
    epochs=1,
    batch_size=32
)

pred_output = model.predict({'input_ids':x_test['input_ids'], 'attention_mask'
    ↪ :x_test['attention_mask']})

predicted_y = np.argmax(pred_output, axis=1)

# Classification report for BERT
print(classification_report(X_test.cyberbullying_type, predicted_y))

# Confusion matrix for BERT
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay

pred = model.predict({'input_ids': x_test['input_ids'], 'attention_mask':
    ↪ x_test['attention_mask']})

class_pred = np.argmax(pred, axis=1)

class_true = np.argmax(to_categorical(X_test.cyberbullying_type), axis=1)

cm_bert = confusion_matrix(class_true, class_pred)

disp = ConfusionMatrixDisplay(confusion_matrix=cm_bert, display_labels=['Class
    ↪ ␣0', 'Class␣1', 'Class␣2', 'Class␣3', 'Class␣4'])
disp.plot(cmap='Blues')
```

Confusion Matrices

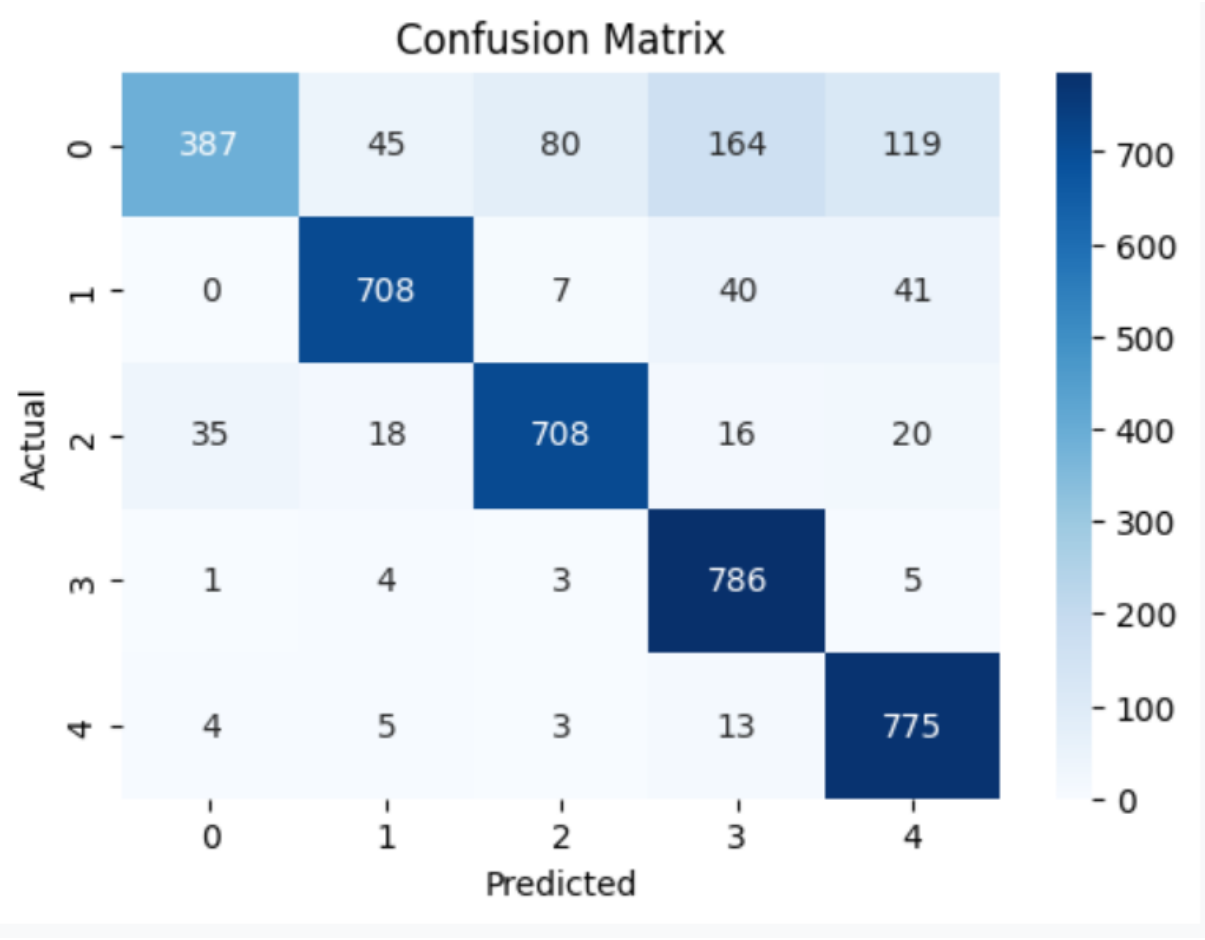


Figure 7: Confusion Matrix for Naive Bayes

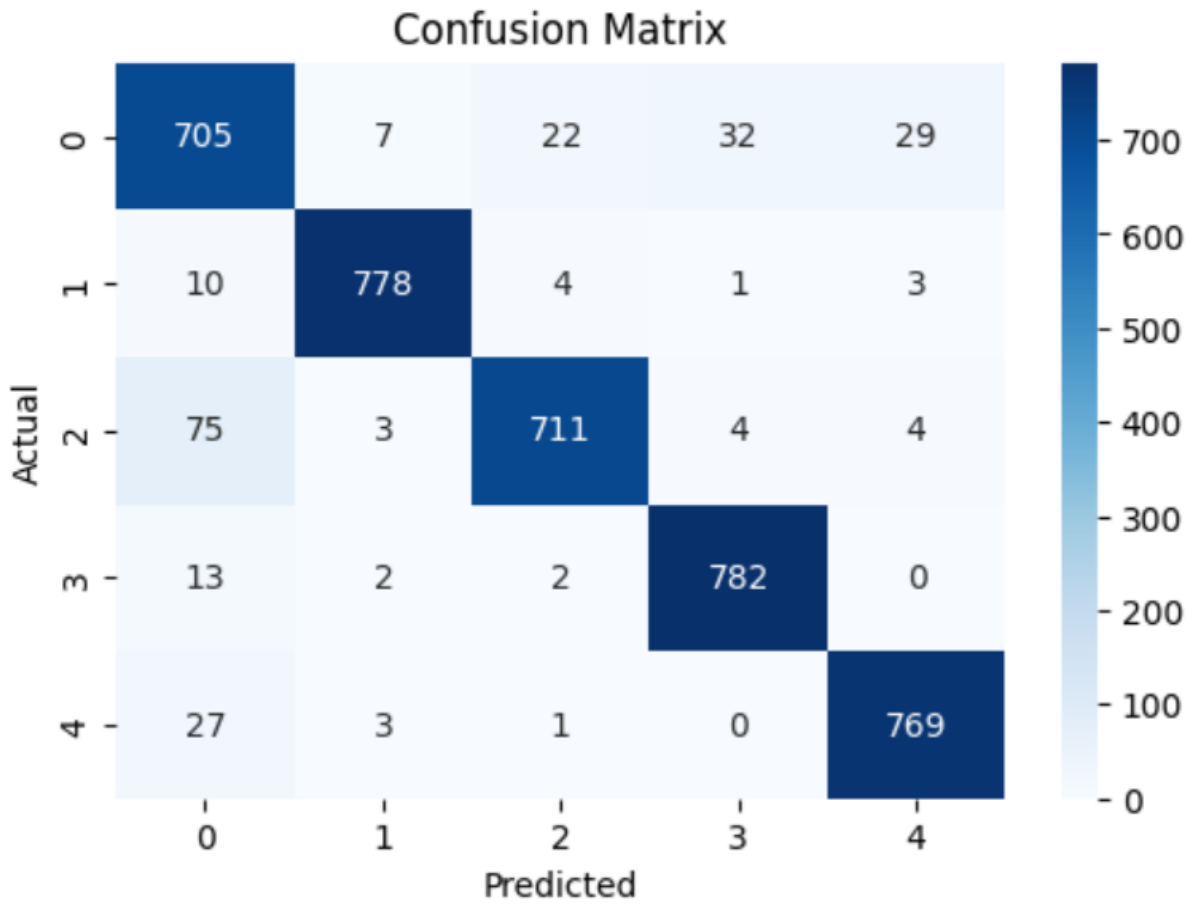


Figure 8: Confusion Matrix for Logistic Regression

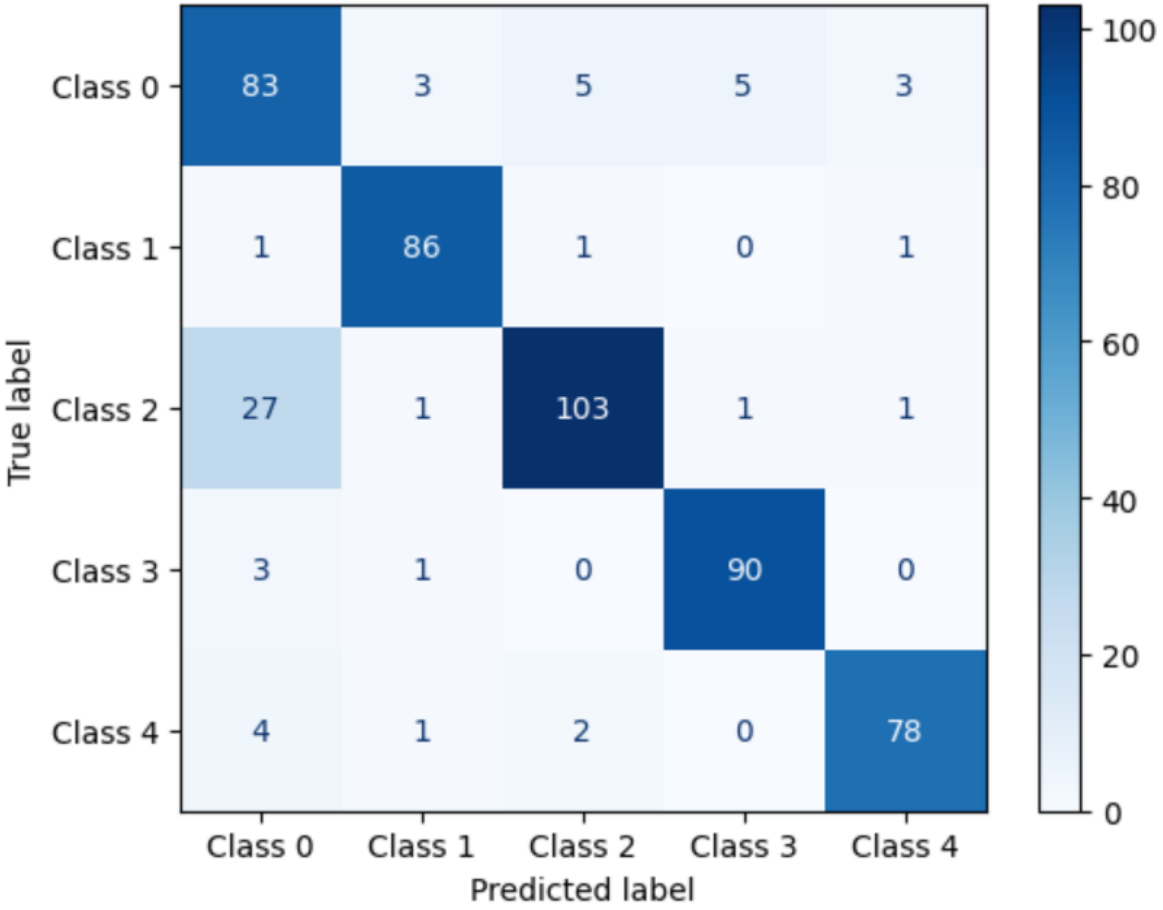


Figure 9: Confusion Matrix for BERT