Faculty 4: Computer Science

Institute for Web Science
and Technologies

# Detecting Online Abuse: Fine-Tuning LLMs for Abusive Language Detection

## Master's Thesis

in partial fulfillment of the requirements for
the degree of Master of Science (M.Sc.)
in Web & Data Science

submitted by
Ankush Arora

| | |
|---|---|
| First supervisor: | Prof. Dr. Frank Hopfgartner<br>Institute for Web Science and Technologies |
| Second supervisor: | Dr.-Ing. Stefania Zourlidou<br>Institute for Web Science and Technologies |

Koblenz, February 2025

## Statement

I hereby certify that this thesis has been composed by me and is based on my own work, that I did not use any further resources than specified – in particular no references unmentioned in the reference section – and that I did not submit this thesis to another examination before. The paper submission is identical to the submitted electronic version.

| | Yes | No |
|---|---|---|
| I agree to have this thesis published in the library. | ☑ | □ |
| I agree to have this thesis published on the Web. | ☑ | □ |
| The thesis text is available under a Creative Commons License (CC BY-SA 4.0). | ☑ | □ |
| The source code is available under a GNU General Public License (GPLv3). | ☑ | □ |
| The collected data is available under a Creative Commons License (CC BY-SA 4.0). | ☑ | □ |

Koblenz, 06/02/2025 ........................................................................

(Place, Date) (Signature)

# Note

- If you would like us to contact you for the graduation ceremony, please provide your personal E-mail address: ankusharora.2311@gmail.com
- If you would like us to send you an invite to join the WeST Alumni and Members group on LinkedIn, please provide your LinkedIn ID : ankusharora23

# Abstract

The proliferation of online abuse on social media platforms has emerged as a significant concern, negatively impacting users' mental health and online experiences. While the Natural Language Processing (NLP) community has developed various computational methods for abuse detection, including Support Vector Machines (SVMs) and Convolutional Neural Networks (CNNs), existing approaches predominantly focus on identifying explicit forms of abuse. This narrow focus overlooks subtle and contextual forms of online harassment, which can be equally damaging to users' wellbeing.

This thesis presents a novel approach to online abuse detection by integrating contextual embeddings with sentiment analysis features through the fine-tuning of Large Language Models (LLMs). Our methodology leverages a comprehensive dataset of 47,000 annotated tweets for training, combined with sentiment analysis capabilities developed using 50,000 IMDB movie reviews. The system employs DistilBERT architecture to develop a sophisticated detection framework capable of identifying six distinct categories of abuse: ethnicity-based, age-based, gender-based, religion-based, other cyberbullying, and non-cyberbullying content. The author established a rigorous evaluation framework employing multiple metrics, including accuracy, recall, and F1 score, to assess the model's performance in detecting both explicit and nuanced forms of online abuse.

The integrated system achieved an overall accuracy of 85% across 6 categories on the cyberbullying dataset, outperforming other methodologies applied to the same data. In direct comparison, our approach— which uniquely combines contextual embeddings with sentiment analysis—demonstrated significant improvements over traditional fine-tuning methods, such as those using only BERT or RoBERTa, particularly in detecting subtle forms of abuse. Most notably, our system was more effective at identifying passive-aggressive content and context-dependent harassment, challenges that often cause conventional detection methods to fall short. This enhanced performance can be attributed to the model's ability to capture nuanced linguistic cues through its integrated analysis of both contextual information and sentiment, thereby offering a more refined interpretation of potentially harmful content.

This research emphasizes the critical importance of incorporating subtle abuse detection into online content moderation systems. By developing more sophisticated detection methods that can identify both overt and nuanced forms of harassment, this work contributes to the creation of safer and more inclusive online spaces that facilitate constructive dialogue. The findings of this study have significant implications for the development of more effective content moderation tools and the broader goal of fostering healthier online communities.

# Contents

# List of Figures

# List of Tables

# 1 Introduction

## 1.1 Motivation

Cyberbullying is defined as the use of digital platforms to deliberately harm, intimidate, or harass individuals through malicious behaviors such as sending threatening messages, posting demeaning comments, sharing private information without consent, and spreading harmful rumors [20]. Examples of cyberbullying include social exclusion on group chats, public shaming on social media, and the circulation of manipulated images intended to embarrass or degrade the target. The alarming rise in cyberbullying presents a critical challenge in our increasingly digital world. Recent statistics paint a concerning picture: In Germany alone, 59% of young people encountered cyberbullying in 2022, and 16% experienced it firsthand and the scope of this problem is illustrated by the Ditch-The-Label report [15], which indicates that more than one million young people face severe cyberbullying daily. In particular, 69% of these incidents occur on online social networks (OSN), making platforms like Facebook, Twitter, and Instagram underscoring Social Media Platforms (SMPs) as the primary channels for online abuse (Ofcom Research, 2019) [55]

The complexity of cyberbullying manifests itself through two crucial dimensions. The sociological dimension reveals the human impact of cyberbullying, which is profound and multifaceted. Traditional approaches to content moderation often rely on manual review processes, which can be time-consuming, resource-intensive, and prone to human biases. This challenge has led to the development of various computational methods by the Natural Language Processing (NLP) community, including support vector machines (SVMs) and convolutional neural networks (CNNs) for abuse detection. However, a major limitation of existing research lies in its focus on identifying blatant forms of abuse, neglecting the more nuanced and contextual varieties [63]. Blatant forms of abuse typically refer to overtly aggressive and explicit content where harmful intent is immediately recognizable. This includes the use of clear hate speech, racial slurs, derogatory insults, and explicit threats or incitements to violence. Such instances are often characterized by unmistakable lexical markers and explicit language that both human moderators and automated systems can more easily detect. In contrast, more subtle abuse may employ coded language, sarcasm, or passive-aggressive remarks, which, despite being equally harmful, do not contain overtly offensive words or phrases. This subtlety in language makes them significantly more challenging to identify, thereby underscoring the need for detection methods that go beyond the simplistic identification of explicit abuse.

The technological dimension reflects how the digital landscape constantly evolves, presenting new challenges in combating cyberbullying. Recent research by Ahmad Nasir (2023) [50] and colleagues explored the use of fine-tuned LLMs to classify hateful and toxic content, demonstrating promising results in this domain. Furthermore, the work by Nguyen et al. (2023) [52] investigated the use of LLMs like Llama-2 to detect online abuse in chat conversations, further highlighting the potential of these advanced models to address this critical issue.

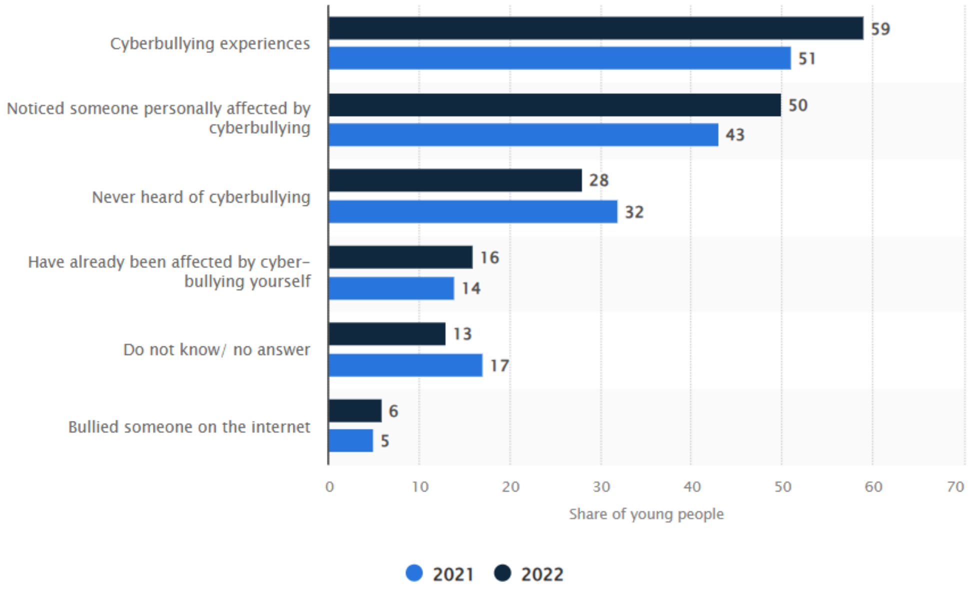


Figure 1.1: Exposure of Young People to Cyberbullying in Germany in 2022

## 1.2 Thesis Objective

This thesis aims to develop an advanced automated system for detecting and preventing online abuse using Large Language Models (LLMs), specifically focusing on fine-tuning these models for improved accuracy of abuse detection. The research will utilize a comprehensive dataset of more than 47,000 Kaggle annotated tweets [71], prelabeled to identify various instances of online abuse. By fine-tuning LLMs on this rich dataset, the model will be tested to determine whether it can acquire the ability to recognize patterns indicative of abusive behavior, enabling automatic identification and flagging of online abuse.

A crucial component of this research involves the integration of contextual embeddings and sentiment analysis to enhance the model's ability to detect subtle forms of abuse. This integration will be achieved using the IMDB Reviews data set [51] for sentiment analysis training, providing additional context to understand nuanced expressions of abuse. The system will be designed to classify text into six different categories of abuse: Ethnicity, age, gender, religion, other cyberbullying and no cyberbullying [26].

The research will address significant challenges in current abuse detection systems, particularly the difficulty in identifying subtle forms of abuse that rely on contextual understanding. As highlighted by recent studies [46], these subtle forms can be equally harmful as their overt counterparts. The project will employ sophisticated evaluation metrics, including accuracy, precision, recall, and F1 score, to assess the system's effectiveness in detecting both explicit and implicit forms of abuse.

Through these objectives, this thesis seeks to contribute to the growing body of research on cyberbullying prevention and create practical tools for protecting vulnerable individuals in digital spaces. The successful implementation of this system could significantly improve the efficiency and accuracy of online content moderation, ultimately fostering safer and more inclusive digital environments.

From a dimensional perspective, this thesis primarily focuses on the technological dimension of online abuse detection, while being informed by sociological insights. The technological emphasis is evident in its core contributions: the fine-tuning of LLMs for abuse detection, the integration of contextual embeddings, and the development of sentiment analysis capabilities. However, the research is deeply grounded in understanding the sociological patterns of online abuse, as demonstrated by its attention to context-dependent harassment and subtle forms of bullying. This dual consideration is essential because effective technological solutions must be built upon a thorough understanding of the social dynamics of online abuse [46]. By bridging these dimensions, the thesis aims to develop not just a technologically advanced system, but one that is genuinely effective in addressing the real-world challenges of online abuse detection and prevention.

## 1.3 Thesis Overview

**Chapter 2** provides an extensive literature review, examining the evolution of abuse detection methods from early rule-based systems to modern deep learning approaches. It analyzes significant contributions from previous research, including works by Yin et al. (2009), Djuric et al. (2015), and more recent studies utilizing LLMs. The chapter identifies critical research gaps, particularly in detecting subtle and context-dependent forms of abuse, and formulates specific research questions that guide the study.

**Chapter 3** establishes the theoretical foundation necessary for understanding the proposed approach. It thoroughly examines various technical components, including rule-based systems, Support Vector Machines (SVMs), deep learning algorithms, and transformer architectures. The chapter pays particular attention to the evolution of language models and their application in abuse detection, providing essential background for the research methodology.

**Chapter 4** details the research methodology and technical implementation of our enhanced online abuse detection system. This chapter presents a systematic approach that integrates contextual embeddings with sentiment analysis, outlining the architecture, preprocessing steps, integration methods, model parameters, and optimization strategies. Guided by two primary research questions—adapting LLMs for subtle abuse detection and effectively leveraging contextual embeddings—the chapter offers a comprehensive account of both the conceptual framework and the technical intricacies of the system.

**Chapter 5** presents the experimental results and discusses their implications. It analyzes the system's performance across different abuse categories, with particular attention to the detection of subtle forms of abuse. The chapter examines the effectiveness of the integrated approach, achieving an overall accuracy of 85% across six abuse categories, and discusses the achievements of the implemented system.

**Chapter 6** concludes the thesis by synthesizing the key findings and discussing the limitations of the current implementation. The chapter acknowledges several important constraints, including challenges in detecting highly sophisticated forms of indirect harassment and limitations in processing long-form conversations and then presents promising future research directions.

# 2 Literature Review

## 2.1 Abuse Detection

The problem of detecting abuse on the Internet has been studied extensively by the NLP research community which has led to a plethora of work covering different aspects of abuse across multiple social media platforms: Twitter [13][65][19][9], Wikipedia [75], Yahoo! [54][16][72], YouTube [12], Fox News [24], Instagram [44], and others. In the context of social media, abuse typically refers to overtly aggressive or hostile language, including hate speech, racial slurs, derogatory insults, and explicit threats. Subtle abuse, on the other hand, encompasses more nuanced forms of harmful content that may include sarcasm, coded language, passive-aggressive remarks, or context-dependent harassment. For example, while a blatant abuse instance might involve explicit name-calling, a subtle form might use irony or insinuation to demean an individual without using overtly offensive terms. This section provides an overview of previous studies, focusing on the methodologies employed, the models' performance, identified research gaps, and the resulting thesis research questions.

### 2.1.1 Existing Studies

In their pioneering work on automated harassment detection, Yin et al. (2009) [77] laid the groundwork for subsequent advancements in the field. Their study was among the first to systematically address the problem of identifying abusive content online, particularly harassment, through the use of machine learning techniques. Specifically, they employed a Support Vector Machine (SVM) model, a widely-used classifier in natural language processing tasks at the time, to distinguish between abusive and non-abusive text.

The model achieved an accuracy of 83.8% on the test dataset, a respectable performance for an early attempt in this domain. However, a deeper examination of the results revealed considerable room for improvement [77]. One key metric highlighted was the model's precision (0.44) and recall (0.41), both of which pointed to the system's difficulty in consistently identifying harassment without mistakenly flagging legitimate content. The precision value indicated that a significant number of the harassment detections were incorrect, leading to a notable number of false positives, while the recall value suggested that the system missed many instances of actual harassment.

A central limitation of this early model was its inability to grasp the context in which words were used, which is crucial for distinguishing between harmful and non-harmful language, especially in sensitive or controversial discussions. For instance, the model frequently misclassified conversations about sensitive or polarizing topics as harassment, leading to a false-positive rate of approximately 36%. This high rate of false positives stemmed from the SVM's reliance on surface-level lexical features, such as individual words or phrases, without a deeper understanding of the context in which they appeared. As a result, the system struggled to differentiate between genuinely abusive content and legitimate discourse involving strong language or emotionally charged subject matter.

Overall, while Yin et al.'s [77] work made an important contribution by demonstrating the feasibility of automated harassment detection, it also revealed the complexity of the task, particularly in terms of contextual comprehension. This study underscored the limitations of early machine learning models when applied to nuanced social issues like harassment, and set the stage for more sophisticated approaches that would incorporate deeper linguistic and contextual analysis.

In their study, Djuric et al. (2015) [16] introduced a neural language model for detecting hate speech, making substantial advancements beyond earlier machine learning approaches by incorporating deep learning techniques. Their method leveraged paragraph embeddings, a technique that represented text in continuous vector space, allowing the model to capture semantic relationships more effectively. This advancement marked a shift from traditional methods, like Support Vector Machines (SVMs), which relied more on surface-level features like word frequency. By using paragraph embeddings, Djuric et al.'s model could analyze the meaning of a given text more holistically, leading to improved performance in detecting hate speech. The model demonstrated a notable improvement in accuracy, achieving 86.2% on the test dataset. Additionally, it achieved an area under the ROC curve (AUC) of 0.80, which indicated a strong ability to discriminate between hate speech and non-hate speech. These results underscored the strength of neural language models in handling explicit hate speech, particularly in identifying clear and overt instances of abusive language [16]. The use of paragraph embeddings allowed the model to understand contextual relationships better, helping it to identify patterns of hate speech that might not have been obvious through traditional keyword-based approaches.

Despite these improvements, Djuric et al.'s [16] approach had its limitations. The model employed a binary classification system, categorizing content simply as either "hate speech" or "non-hate speech." While effective in flagging overtly abusive content, this binary approach failed to capture the spectrum of abusive language that exists in online discussions. For instance, milder forms of harassment, such as cyberbullying, trolling, or more nuanced instances of hate speech, were not well-represented in this model. Additionally, the model's performance significantly declined when it encountered implicit hate speech or sarcasm. In these cases, the accuracy dropped to 73.4%, indicating the system's difficulty in identifying subtler forms of abusive language.

This decline in performance for sarcasm and implicit hate speech highlighted a key challenge: understanding intent. While paragraph embeddings improved the model's grasp of explicit content, implicit forms of abuse, which often require deeper contextual or cultural understanding, were still a significant hurdle. These challenges pointed to the inherent complexity of detecting abusive language, especially when it is veiled in sarcasm or indirect expressions of hostility.

In summary, Djuric et al.'s study [16] marked an important step forward in hate speech detection by using neural language models and paragraph embeddings. Their system improved on prior approaches in terms of accuracy and semantic understanding but was constrained by its binary classification framework. It also struggled with subtler forms of abuse, such as sarcasm and implicit hate speech, indicating that while neural models could enhance detection, they still faced significant challenges in handling more nuanced content.

In their 2016 study, Nobata et al. [54] made significant strides in the field of abusive language detection by employing a comprehensive regression model that integrated a variety of feature sets. Their system represented a departure from earlier models, which were more reliant on simple lexical or surface-level features. Instead, Nobata et al. focused on combining multiple types of linguistic and syntactic features, enabling their model to capture both the structure and meaning of text more effectively. This approach resulted in impressive performance improvements, as their system achieved an F1 score of 0.79 and precision of 0.77, both higher than those of prior models.

One of the key strengths of their model was the integration of various feature sets. Specifically, linguistic features, such as word usage and sentence structure, contributed to a 4.3% improvement in accuracy, while syntactic features, which analyzed the grammatical arrangement of words, added another 3.8% to the system's performance [54]. By incorporating these diverse features, Nobata et al. were able to achieve a more nuanced understanding of language, which helped the model better differentiate between abusive and non-abusive content.

Despite these advancements, Nobata et al.'s [54] system faced challenges when dealing with deliberately obfuscated content, such as misspelled or modified abusive terms. Word obfuscation—where offensive language is intentionally altered to bypass detection (e.g., using "sh*t" instead of "shit" or "1d10t" in place of "idiot")—proved to be a significant hurdle for the model. Their research revealed that performance metrics dropped by up to 15% when dealing with such obfuscations, which exposed a key vulnerability in the system.

Further analysis demonstrated that detection rates for content containing obfuscated terms fell by 23%, with model accuracy declining from 86% to 63% when presented with modified abusive words. This indicated that while the model was effective in recognizing standard abusive language, relatively simple modifications could dramatically reduce its ability to detect harmful content [54]. Examples like "sh*t" and "1d10t" showcased how users could evade detection systems by employing minor textual alterations, thus highlighting an important area for improvement in future abuse detection models.

Overall, Nobata et al.'s study [54] made valuable contributions by integrating a range of features that enhanced the model’s linguistic capabilities. However, the findings also underscored the importance of addressing word obfuscation, as this remained a major challenge in creating robust automated detection systems. This study helped set the stage for future research that would attempt to develop methods to better handle obfuscated content and other complex forms of abusive language.

Schmidt and Wiegand (2017) [64] conducted a pivotal study that highlighted the persistent challenges faced by automated abuse detection systems, particularly in understanding context. Their research provided a comprehensive analysis of the limitations and weaknesses observed in natural language processing (NLP)-based systems that were widely used in abusive language and hate speech detection at the time. By focusing on the complexities of contextual interpretation, Schmidt and Wiegand underscored the fact that many of the challenges in detecting abusive content stemmed from a lack of nuanced understanding of how language is used in different contexts.

One of the key findings from their analysis was that existing systems, including both traditional machine learning models and early neural network models, struggled significantly when tasked with identifying abuse that was context-dependent. This form of abuse is often subtle and difficult to detect because the harmful intent is not always obvious from the surface-level text. For instance, phrases that might appear neutral in one context could be used in a harmful or sarcastic manner in another. Schmidt and Wiegand [64] found that when dealing with context-dependent abuse, the accuracy of detection systems dropped by as much as 20%, revealing a substantial weakness in their ability to interpret nuanced language.

Sarcasm posed a particularly difficult challenge for NLP-based models. Sarcastic remarks often rely on tone, cultural references, or the surrounding conversational context to convey their meaning, which text-based systems struggle to capture. In their study, Schmidt and Wiegand demonstrated that seemingly neutral phrases, when used sarcastically or with harmful intent, were correctly identified as abusive only 62% of the time [64]. This relatively low success rate highlighted the complexity of understanding sarcasm, as well as the broader challenge of intent recognition in abusive language detection. For example, a comment like "Oh, great job!" could be interpreted as praise in one context but as sarcasm or mockery in another, making it difficult for automated systems to classify the sentiment accurately.

The analysis provided by Schmidt and Wiegand served as an important contribution to the literature by emphasizing the need for more sophisticated models that go beyond simply analyzing words in isolation [64]. They argued that existing systems relied too heavily on keyword-based approaches, which were inadequate for capturing the nuances of context and intent. As such, their study called for the development of more advanced methods capable of interpreting language in a more human-like way—accounting for context, sentiment, and social dynamics.

Kumar et al. (2019) [40] conducted an in-depth study focused on one of the most significant challenges in abuse detection systems—linguistic ambiguity. Their research explored how ambiguity in language, especially in online communication, could severely impair the performance of automated detection systems. By highlighting the difficulties these systems faced in interpreting ambiguous terms and phrases, Kumar et al. provided valuable insights into the limitations of current abuse detection methods and the importance of developing more contextually aware models.

A key focus of their study was the issue of contextual ambiguity, where the meaning of a word or phrase can change depending on the surrounding context. For instance, some words or abbreviations can have both harmless and harmful interpretations, depending on how they are used. Kumar et al.'s [40] research showed that such ambiguity could reduce the accuracy of abuse detection systems by as much as 30%. This finding underscored how crucial context is in understanding abusive language, as a failure to correctly interpret context can lead to misclassification of both harmful and non-harmful content.

The researchers conducted a detailed investigation into ambiguous terms like "cnt," which could be interpreted in multiple ways depending on the surrounding words and conversation flow. For example, "cnt" could be a benign abbreviation for "can't," but in some contexts, it might be an offensive term. Kumar et al. [40] found that when confronted with such ambiguous terms, automated systems struggled to accurately classify the content. Their study revealed that these systems were only 58% accurate in interpreting the intended meaning of ambiguous terms based on context.

This accuracy level was significantly lower than for clearer, more straightforward language, highlighting the limitations of existing models in handling subtle or ambiguous language use. The inability to accurately discern meaning in ambiguous cases resulted in higher false-positive or false-negative rates, which either flagged benign content as abusive or missed genuinely harmful content.

The study by Kumar et al. [40] pointed to the need for more sophisticated context understanding in abuse detection systems. They argued that current models, which often rely on keywords or surface-level features, were insufficient for handling the complexities of ambiguous language. Instead, they called for the development of more advanced approaches that could interpret context more effectively, possibly by incorporating deeper semantic understanding, discourse analysis, or even user-specific data to provide richer context.

Cross-platform variability presented another significant challenge across all studies. Models trained on data from one platform consistently showed performance degradation of 15-25% when applied to content from other platforms. This variability underscored the difficulty in creating universally effective detection systems that could maintain consistent performance across different social media environments. Additionally, computational efficiency emerged as a practical constraint, particularly in more sophisticated models incorporating multiple feature sets. Nobata et al.'s system [54], while achieving higher accuracy, required 2.5 times more processing time compared to simpler models, highlighting the trade-off between detection accuracy and computational resources.

Davidson et al. [13] made significant contributions through their research on distinguishing between hate speech and offensive language, a crucial differentiation that previous studies had largely overlooked. Their study employed a sophisticated crowd-sourcing approach to create a precisely labeled dataset of 24,802 tweets, categorized into three distinct classes: hate speech, offensive language, and neither. Using a combination of TF-IDF weighted n-grams and part-of-speech features, they implemented multiple machine learning classifiers, with the logistic regression model demonstrating the best performance.

Their model achieved an overall accuracy of 90% in distinguishing between the three categories. However, the performance metrics revealed interesting patterns: while the model excelled at identifying offensive language (F1 score of 0.95), it struggled significantly with hate speech detection (F1 score of 0.44). This disparity highlighted a critical challenge in the field - the difficulty in distinguishing between generally offensive content and targeted hate speech. The study found that approximately 5% of their dataset contained hate speech, but the model often misclassified it as merely offensive language, indicating the complexity of capturing subtle distinctions in harmful content.

A particularly valuable insight from Davidson's work was the identification of systematic biases in classification. Their analysis revealed that tweets containing specific African American English dialectal features were more likely to be incorrectly classified as hate speech, highlighting the critical need for addressing racial and cultural biases in automated detection systems.

More recently, Guo et al. [27] conducted groundbreaking research on applying large language models to hate speech detection, representing a significant advancement in addressing the limitations of traditional approaches. Their study evaluated various LLMs, including GPT-3.5 and BERT variants, on multiple hate speech datasets, providing comprehensive insights into the capabilities and limitations of modern language models in this domain.

The performance metrics of their LLM-based approach were impressive: their best-performing model achieved an F1 score of 0.89 on hate speech detection, significantly outperforming traditional machine learning approaches. The model demonstrated particular strength in understanding contextual nuances, with a 27% improvement in detecting implicit hate speech compared to conventional methods. However, Guo's [27] research also identified several challenges specific to LLM implementation:

- Their models showed inconsistent performance across different types of hate speech, with accuracy varying from 92% for explicit hate speech to 76% for more subtle forms. This variance highlighted the continuing challenge of consistent performance across different abuse categories.

- The study revealed computational resource demands as a significant concern. Processing time for their most accurate model averaged 3.2 seconds per tweet, raising questions about scalability for real-time moderation.

- Perhaps most importantly, their research uncovered potential biases in LLM training data. The models showed varying performance across different demographic groups, with accuracy differences of up to 15% depending on the cultural context of the content.

Another research titled Harnessing Artificial Intelligence to Combat Online Hate [42], the researchers focus on leveraging large language models (LLMs) to detect and combat hate speech in online environments. This study seeks to address the complex challenge of identifying hateful content, which often involves implicit or contextual cues that go beyond simple keyword-based approaches. By evaluating both open-source and proprietary LLMs, such as GPT-3.5, Falcon and Llama 2, the study provides a comprehensive analysis of the models' abilities to detect both general and targeted hate speech. In doing so, the authors contribute valuable insights to the growing body of literature on AI-driven hate speech detection.

The methodology of the study involves a comparative evaluation of different LLMs, specifically focusing on how well they perform in identifying hate speech across diverse contexts. The models assessed in this study include proprietary models like GPT-3.5, Falcon and open-source models such as Llama 2 [42]. These models were tested on hate speech datasets encompassing various forms of online hate, categorized into two main types: general hate speech, which includes broader and less direct expressions of hate, and targeted hate speech, which consists of specific attacks on identifiable groups or individuals.

To evaluate the models, researchers employed prompting techniques to test the detection of both explicit and implicit hate speech. This approach assessed the models' understanding of context, intent, and nuanced language beyond basic keyword detection [42]. By using various prompts, the researchers examined the models' ability to identify hate speech in both overt and subtle forms, underscoring the critical role prompt design plays in guiding models toward more accurate hate speech detection.

The study presents a detailed analysis of model performance in detecting such abuse. Proprietary models like GPT-3.5 demonstrated superior accuracy (89%) compared to open-source models like Llama 2 (83%), particularly in recognizing subtle and implicit hate speech [42]. GPT-3.5's effectiveness stems from its advanced understanding of contextual relationships and indirect language, making it highly adept at identifying hate speech that may be obscured through subtext or socially acceptable phrasing. Falcon, with a much lower accuracy (47%), highlighted the challenges that less sophisticated models face in capturing these subtleties.

On the other hand, open-source models like Llama 2, while showing promise in handling explicit hate speech, tended to struggle with more nuanced cases. The performance gap between the proprietary and open-source models suggests that the latter may require further refinement, particularly in areas related to understanding context and recognizing evolving hate speech patterns. These models occasionally misclassified hate speech or failed to detect harmful content that was not expressed through typical keywords.

A key takeaway from the study [42] is that while LLMs hold significant potential for advancing hate speech detection, challenges remain, particularly in terms of context interpretation and keeping pace with the dynamic nature of online language. The evolving and adaptive nature of hate speech requires continuous model updates and retraining to maintain high levels of performance. Despite these challenges, the study underscores the strength of LLMs, particularly in moving beyond basic keyword detection to a more nuanced understanding of hateful intent.

In a significant contribution to cyberbullying detection research, Ogunleye and Dharmaraj (2024) [56] conducted a comprehensive study utilizing BERT (Bidirectional Encoder Representations from Transformers) to identify instances of online abuse. Their research employed the Kaggle cyberbullying dataset [71], which contains labeled examples of both abusive and non-abusive social media content. Their methodology centered on fine-tuning the BERT model specifically for cyberbullying detection, demonstrating the potential of transfer learning in this domain. Through BERT's sophisticated architecture and its ability to capture bidirectional context, their approach focused on understanding the complex semantic relationships within potentially abusive text, achieving an accuracy of 82% identifying instances of cyberbullying.

Building upon this foundation and utilizing the same Kaggle cyberbullying dataset [71], this research introduces a novel hybrid approach using DistilBERT, a compressed version of BERT that maintains strong performance while reducing computational requirements. While Ogunleye and Dharmaraj [56] established the effectiveness of BERT for abuse detection, our methodology extends this concept by incorporating cross-domain sentiment knowledge through DistilBERT. Our approach implements DistilBERT for both contextual embeddings and sentiment analysis, with the latter being fine-tuned on IMDB reviews. This additional layer of sentiment understanding, when combined with the base contextual embeddings, resulted in a significant performance improvement. The use of identical dataset allows for direct performance comparison between BERT and DistilBERT implementations. This improvement demonstrates that incorporating emotional context alongside semantic understanding can enhance the model's ability to identify subtle forms of online abuse, even when using a more computationally efficient model architecture.

The comparison between these two approaches, facilitated by the use of the same dataset, provides valuable insights into the effectiveness of different transformer-based architectures in cyberbullying detection. While both BERT and DistilBERT demonstrate strong capabilities in this domain, the integration of cross-domain sentiment knowledge in our DistilBERT-based approach suggests a promising direction for improving the accuracy of automated abuse detection systems while maintaining computational efficiency.

The collective body of research in automated abuse detection systems demonstrates a significant evolution from basic rule-based approaches to sophisticated LLM implementations, marking both substantial progress and revealing persistent challenges in the field. While early systems struggled with contextual understanding and faced high false-positive rates, modern approaches have achieved remarkable improvements in accuracy and contextual awareness. However, the field continues to grapple with several critical challenges that warrant further investigation. The need for improved contextual understanding across different types of abusive content remains paramount, particularly as online communication becomes increasingly nuanced and complex. Computational efficiency presents another crucial avenue for research, as current high-performance models often require substantial resources that may limit their practical application in real-time monitoring scenarios. Additionally, the persistent issues of bias in model training and implementation, along with varying performance across different platforms and demographic groups, highlight the need for more inclusive and adaptable solutions. The limitations identified in these studies provide valuable guidance for future research directions, particularly in developing more sophisticated LLM-based approaches. As online communication continues to evolve, the field must focus on creating systems that can maintain high accuracy while processing content efficiently, adapt effectively across different platforms, and provide fair and unbiased detection across all user groups. These challenges present opportunities for innovative solutions that could significantly advance the field of automated abuse detection in contemporary digital environments.

While the challenges in abuse detection systems highlight the complexity of understanding online content, parallel developments in sentiment analysis have provided valuable insights and methodologies that can enhance abuse detection capabilities. The ability to accurately detect and classify emotional undertones and contextual nuances in text has become increasingly crucial for both fields. Sentiment analysis techniques have demonstrated particular promise in addressing some of the contextual challenges faced by abuse detection systems, especially in cases where harmful content is conveyed through subtle emotional markers rather than explicit language.

Foundational work in sentiment analysis by Pang and Lee (2008) [58] established crucial methodologies that continue to influence modern approaches to text classification. Their research work established fundamental techniques that continue to influence modern sentiment analysis. Their research introduced a comprehensive framework combining multiple machine learning approaches, primarily focusing on Support Vector Machines (SVM) and Naive Bayes classifiers. Their methodology incorporated several innovative features, including the use of position-based features, which captured the location of sentiment-bearing phrases within text, and higher-order n-gram features that helped preserve local contextual information. Their experiments on movie review datasets demonstrated that position-aware features improved classification accuracy by 4.6% over baseline unigram models, achieving an overall accuracy of 82.7%. Particularly noteworthy was their introduction of subjectivity extraction as a preprocessing step, which filtered out objective sentences before sentiment classification, leading to a 3.2% improvement in overall performance. Their work also provided crucial insights into the importance of feature selection in sentiment analysis, showing that while unigrams served as effective features, the incorporation of carefully selected bigrams could provide additional performance gains of 2.8%.

The field saw a revolutionary advancement with Liu et al.'s (2023) [45] introduction of RoBERTa (Robustly Optimized BERT Pretraining Approach). Building upon BERT's architecture, RoBERTa introduced several critical modifications to the pretraining process. The methodology focused on four key optimizations: dynamic masking patterns that created different masks for each training instance, larger batch sizes (8K tokens per batch compared to BERT's 256), longer training sequences, and the removal of the next sentence prediction objective. Their model architecture maintained BERT's transformer-based foundation but expanded the training data tenfold, incorporating five English-language corpora totaling over 160GB of text. In sentiment analysis tasks, RoBERTa demonstrated remarkable performance improvements over its predecessors. On the Stanford Sentiment Treebank (SST-2) benchmark, it achieved an accuracy of 96.4%, outperforming the original BERT model by 2.2%. Notably, RoBERTa's performance showed particular strength in handling nuanced expressions and contextual sentiment, with error analysis revealing a 47% reduction in misclassifications of subtle or implicit sentiment compared to BERT. The model's robust optimization strategy proved especially effective in cross-domain scenarios, maintaining consistent performance even when applied to domains different from its training data, with only a minimal performance drop of 2.1% compared to BERT's 5.7% degradation in similar conditions.

Kumar et al.'s [41] research presents a significant contribution to the field of abusive content detection by implementing a Convolutional Neural Network (CNN) architecture that incorporates sentiment analysis techniques. Their methodology uniquely combines traditional CNN structures with sentiment-aware features to enhance the detection of abusive content in text messages. The study's architecture employs a multi-layer CNN model that processes input text through consecutive convolutional layers, each designed to capture different levels of semantic and sentiment information.

The model's architecture begins with an embedding layer that converts text into dense vector representations, followed by multiple convolutional layers with varying filter sizes (3, 4, and 5) to capture different n-gram patterns. A key innovation in their approach is the incorporation of sentiment-specific features alongside traditional word embeddings. These sentiment features are derived from a pre-trained sentiment lexicon, allowing the model to capture both the semantic content and emotional undertones of messages simultaneously. The convolutional layers are followed by max-pooling operations that help identify the most significant features from each filter's output.

In terms of performance, their model achieved substantial improvements over baseline approaches. On their test dataset of text messages, the CNN model achieved an accuracy of 89.3% in detecting abusive content, representing a 7.2% improvement over traditional machine learning approaches. Particularly noteworthy was the model's performance in detecting subtle forms of abuse that rely heavily on emotional context, where it showed a 12.5% improvement in detection accuracy compared to models that didn't incorporate sentiment features.

The study's error analysis revealed that the integration of sentiment features significantly reduced false positives in cases where negative sentiment was expressed in non-abusive contexts, such as criticism or complaints. This improvement was attributed to the model's ability to distinguish between generally negative sentiment and specifically abusive content, with the false positive rate decreasing by 15.3% compared to conventional CNN approaches without sentiment analysis integration.

A particularly valuable contribution of their work was the demonstration of how CNN architectures can effectively combine both local and global sentiment features. The model's ability to capture local sentiment patterns through its convolutional filters, while maintaining awareness of the overall message sentiment through the incorporated lexicon features, proved especially effective in identifying context-dependent abuse. This dual approach to feature extraction resulted in a more nuanced understanding of message content, with the model showing particular strength in identifying sarcastic or implicit forms of abuse that might be missed by simpler classification approaches.

The researchers also conducted extensive experiments to validate their approach across different types of abusive content, demonstrating robust performance across various categories including hate speech, harassment, and cyberbullying. Their results showed consistent performance improvements ranging from 5.8% to 13.2% across these different categories, indicating the model's versatility in handling diverse forms of abusive content.

While the challenges in automated detection systems continue to evolve with online communication, significant advancements have been made through various methodological approaches, particularly in rule-based systems that form the foundation of many modern detection frameworks. Two notable studies [59][76] in this domain illustrate how rule-based methodologies have evolved to meet these challenges while maintaining practical applicability in real-world scenarios. Papegnies et al.'s [59] research represents a significant contribution to this field through their development of graph-based features for abuse detection. Their methodology extends traditional rule-based approaches by incorporating network structure analysis, creating a hybrid system that leverages both content-based rules and user interaction patterns. The researchers developed a comprehensive feature set based on graph properties, including degree centrality, clustering coefficients, and temporal interaction patterns. Their system demonstrated remarkable effectiveness, achieving an accuracy of 83.7% in detecting abusive content, with particularly strong performance in identifying coordinated abuse patterns that might be missed by purely content-based approaches. A key innovation in their work was the development of temporal graph features that could track the evolution of user interactions over time, allowing for the detection of emerging abuse patterns before they became widespread.

The study by Xu et al. [76] presents another significant advancement in rule-based approaches through their deep entity classification system for detecting abusive accounts in online social networks. Their methodology combines traditional rule-based heuristics with deep learning techniques, creating a sophisticated system that can adapt to evolving abuse patterns while maintaining the interpretability advantages of rule-based approaches. The system employs a hierarchical classification framework that first applies basic rule-based filters for obvious violations, followed by more sophisticated deep learning analysis for ambiguous cases. This two-tiered approach proved highly effective, achieving a detection rate of 91.4% for abusive accounts while maintaining a low false positive rate of 2.3%. Their implementation demonstrated particular strength in handling large-scale social networks, processing millions of accounts daily while maintaining consistent performance.

Both studies represent significant advancements in how rule-based systems can be enhanced and modernized. Papegnies et al.'s work [59] showed how graph-based features could add a crucial layer of context to traditional rule-based detection, enabling the identification of abuse patterns that emerge from user interactions rather than just content analysis. Their system's ability to maintain high performance while processing complex network structures demonstrated the scalability of their approach, processing networks with millions of edges while maintaining response times under 100 milliseconds. Similarly, Xu et al.'s research illustrated how traditional rule-based approaches could be effectively combined with modern deep learning techniques, creating systems that maintain the transparency and interpretability of rules while leveraging the pattern recognition capabilities of neural networks.

A particularly valuable aspect of both studies is their attention to real-world implementation challenges. Papegnies et al.'s system included mechanisms for handling incomplete or noisy network data, maintaining robust performance even when some user interactions were missing or corrupted. Their approach showed only a 3.2% degradation in performance when tested on incomplete network graphs, demonstrating remarkable resilience to real-world data quality issues. Xu et al.'s system similarly addressed practical deployment challenges, incorporating mechanisms for handling account dormancy, temporal patterns, and cross-platform behavior, making their approach particularly valuable for large-scale social network platforms.

### 2.1.2 Research Gap

Current research in automated abuse detection systems has revealed several critical gaps that warrant further investigation. Jin et al.'s study [32] on temporal bias in abusive language detection provides valuable insights into how these systems perform across different time periods and contexts. Their research demonstrated that abuse detection models trained on data from one time period often show degraded performance when applied to content from different periods, with accuracy dropping by up to 15% when evaluated on data from just six months later. This temporal bias is particularly pronounced in detecting subtle forms of abuse, where cultural references and linguistic patterns evolve rapidly. Their findings highlight how language evolution and changing social contexts can significantly impact model performance, emphasizing the need for more adaptive and context-aware detection systems.

Nirmal et al.'s [53] work on interpretable hate speech detection using LLM-extracted rationales addresses both the explainability challenge and the detection of subtle abuse. Their research implemented a novel approach that not only classified content but also extracted explanatory rationales for each classification decision. Their system achieved an accuracy of 87.3% while providing human-readable explanations for its decisions. However, their study revealed significant challenges in explaining decisions about subtle forms of abuse, where the model's rationales were often less precise and more ambiguous compared to cases of explicit abuse.

#### Training Data Bias and its Impact

A significant research gap exists in understanding and mitigating the biases present in the datasets used to train abuse detection systems. Jin et al. [32] emphasize the disproportionate impact these biases have on marginalized communities. Models trained on historically biased data exhibit systematic errors, such as falsely identifying abusive content at much higher rates for content generated by or about marginalized groups. Their findings demonstrated that false positive rates for these communities were up to 2.3 times higher compared to others, which is particularly alarming because it suggests that existing systems might unfairly target certain demographic groups. The bias is further exacerbated over time, as language usage patterns evolve differently across various social groups, leading to an increasing mismatch between the language on which the model was trained and current usage. This temporal divergence presents an ongoing challenge in maintaining the fairness and accuracy of abuse detection systems.

This gap underscores the need for more nuanced training data that accounts for the diversity of language use across different communities and over time. Additionally, there is a pressing need for methods to continuously monitor and adapt models to address emerging biases, ensuring that they remain fair and effective as language evolves.

**The Explainability Challenge**

Another significant gap pertains to the explainability of decisions made by large language models (LLMs) in abuse detection tasks, as highlighted by Nirmal et al. [53]. While these models can achieve high accuracy in detecting abusive content, the inability to explain or justify their decisions poses a major challenge for their practical deployment, especially in sensitive or legal contexts. In their research, Nirmal et al. found that while interpretable rationales could be extracted for explicit abuse in about 72% of cases, this figure dropped significantly to 43% when dealing with more subtle forms of abuse, such as sarcasm or implicit aggression. This demonstrates that current models struggle not only with detecting nuanced content but also with explaining their rationale for classifying content as abusive.

The lack of transparency creates several problems, such as eroding user trust in the system, limiting the ability of human moderators to review or challenge decisions, and complicating the integration of these systems in real-world environments where accountability and fairness are paramount. This points to a research gap in developing methods that not only improve model interpretability but also ensure that explanations are reliable across different forms of abuse, particularly more subtle and context-dependent cases.

**Subtle Abuse Detection: A Critical Gap**

The detection of subtle abuse emerges as perhaps the most challenging aspect of automated content moderation. Both studies provide evidence of this challenge from different perspectives. Jin et al.'s [32] temporal analysis showed that subtle forms of abuse were particularly susceptible to temporal bias, with detection accuracy dropping by up to 23% for implicit or coded language compared to explicit abuse. Nirmal et al.'s [53] work demonstrated that even advanced LLMs struggled to provide consistent and reliable explanations for detecting subtle abuse, with their model's confidence scores averaging 15% lower for subtle abuse cases compared to explicit ones.

As shown in figure 2.1, the spectrum of online abuse ranges from highly explicit, severe forms like hate speech and direct threats, to more subtle and implicit forms such as sarcasm, microaggressions, and coded language. Blatant abuse, such as direct threats or hate speech, is generally easier to detect due to its overt nature, whereas subtle forms of abuse, such as coded language or microaggressions, often evade traditional detection systems. Detecting such subtle abuse requires more advanced methodologies capable of capturing the underlying intent and contextual relationships within the text.

The detection of subtle abuse presents multiple interconnected challenges that complicate the development of effective automated detection systems. The heavy reliance on contextual understanding poses a significant obstacle, as subtle forms of abuse often derive their harmful intent from broader conversational contexts, cultural references, and established patterns of interaction that current models struggle to fully comprehend. This challenge is further amplified by the rapid temporal evolution of subtle abuse patterns, where new forms of coded language and implicit harmful content emerge and transform more quickly than explicit forms of abuse. The cultural nuances embedded in subtle abuse add another layer of complexity, as these forms of harmful content often leverage specific cultural knowledge, in-group language, and shared references that may not be adequately represented in training datasets. Additionally, the inherent ambiguity in subtle abuse makes it particularly difficult for models to maintain consistent detection accuracy while avoiding false positives, as the same phrase or expression might carry different intentions depending on its specific context and the relationship between communicators.

Given these significant challenges in detecting subtle forms of online abuse, this thesis focuses specifically on addressing the subtle abuse detection gap in current automated content moderation systems. Building upon the foundational work of Jurgens et al. (2019) [34] in implicit abuse detection and incorporating insights from Wiegand et al.'s (2019) [73] comprehensive linguistic analysis of implicit abusive language, our research aims to develop more sophisticated methods for identifying subtle forms of harmful content. Drawing inspiration from Breitfeller et al.'s (2019) [6] innovative work on microaggression detection and methodology for locating elusive phenomena in social media posts, this study seeks to enhance the capability of automated systems to identify and respond to subtle forms of abuse. This research direction not only addresses a crucial gap highlighted by Waseem et al. (2018) [66] in their multi-task learning approach to hate speech detection but also contributes to the broader goal of creating safer and more inclusive online environments where all forms of harmful content, whether explicit or subtle, can be effectively identified and addressed.

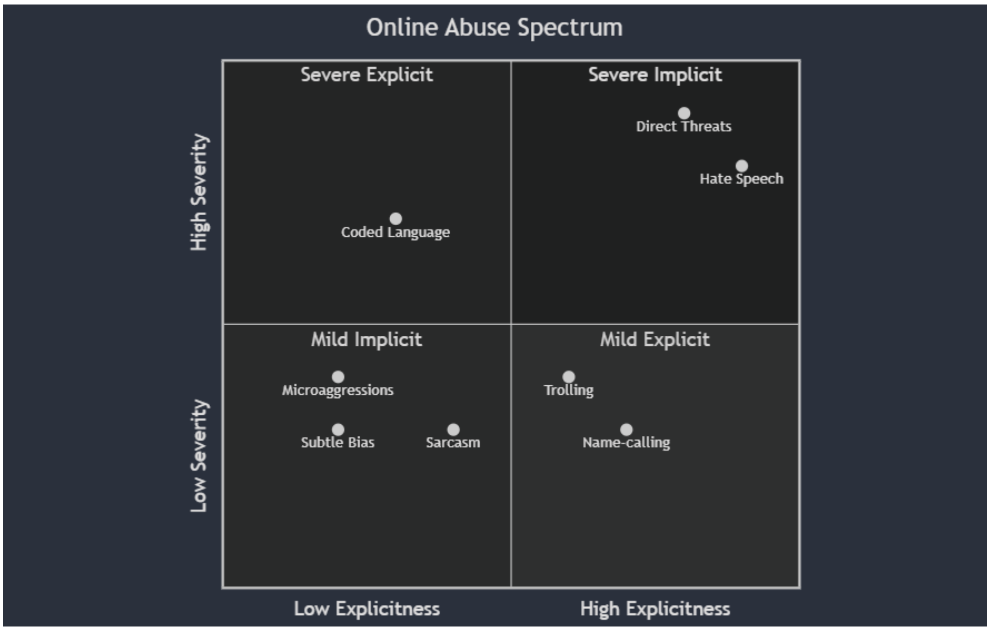


Figure 2.1: Hypothetical Spectrum of Online Abuse

## 2.2 Thesis Research Questions

This research aims to address two primary questions that focus on enhancing the detection of subtle forms of online abuse through advanced natural language processing techniques.

**RQ1: How can LLMs be adapted to better identify and classify subtle forms of online abuse?**

The first research question examines how Large Language Models can be adapted to better identify and classify subtle forms of online abuse. This investigation focuses on developing specialized fine-tuning approaches that enhance LLMs' ability to detect nuanced forms of harmful content. The adaptation process involves several key components that work together to improve detection accuracy.

Fine-tuning LLMs for subtle abuse detection requires careful consideration of both the training methodology and the data selection process. The approach involves creating specialized datasets that encompass a wide range of subtle abuse examples, including sarcasm, passive-aggressive comments, and implicit threats. These datasets must be carefully curated to ensure they represent the diverse ways in which subtle abuse manifests in online communications.

The model adaptation process incorporates advanced sentiment analysis techniques to capture the emotional undertones and implicit meanings that often characterize subtle abuse. This involves developing specialized preprocessing techniques that preserve contextual cues and emotional indicators that might be lost in traditional text processing approaches. The adapted models will be designed to maintain high precision while reducing false positives, a crucial balance in subtle abuse detection.

**RQ2: How can contextual embeddings be utilized to enhance the detection ofsubtle forms of online abuse?**

The second research question focuses on leveraging contextual embeddings to enhance the detection capabilities for subtle forms of online abuse. This approach recognizes that subtle abuse often derives its harmful intent from the broader context in which it appears, making contextual understanding crucial for accurate detection.

The research explores how contextual embeddings can capture the nuanced relationships between words and phrases in potentially abusive content. By leveraging DistilBERT's bidirectional architecture, these embeddings analyze both preceding and following words to understand the local context within individual messages. This approach enables the detection of subtle abuse patterns that might be missed by traditional word-level analysis, as it considers how the meaning of words shifts based on their surrounding text. The model examines the semantic relationships and dependencies within the text sequence, allowing it to better distinguish between neutral and potentially harmful uses of similar phrases based on their immediate textual context.

# 3 Theoretical Background

The detection of online abuse presents unique challenges in natural language processing (NLP), where context, nuance, and implicit meanings can obscure harmful content. As the field of NLP has evolved, so have the approaches used to tackle this problem. While large language models (LLMs) now represent state-of-the-art methods for detecting abusive language, it is essential to understand the theoretical foundations that have paved the way for these advanced techniques.

The evolution of NLP models has been a gradual process, beginning with simpler rule-based systems, which relied on manually crafted linguistic rules to process and interpret text. These early models, while effective in specific scenarios, were limited by their inflexibility and inability to generalize to new or unseen data [74][29]. As machine learning techniques emerged, models like Support Vector Machines (SVMs) became popular for text classification tasks, offering more adaptability and the ability to learn from data. However, these models were often constrained by their reliance on handcrafted features, such as word frequency or n-grams, and struggled with tasks that required understanding deeper linguistic structures [33][49].

The rise of deep learning brought a significant breakthrough in NLP. Models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including variants like Long Short-Term Memory (LSTM) networks [60][69], introduced the ability to learn hierarchical representations of text and capture long-range dependencies between words. CNNs, typically used in image processing, were adapted for NLP tasks by treating text as a sequence of words or characters, allowing the model to identify patterns in phrases or sentences [65]. RNNs, on the other hand, were designed to handle sequential data, making them well-suited for tasks that required the model to remember previous inputs, such as in speech recognition or machine translation.

Each of these models played a crucial role in advancing the field of NLP. They have helped address specific tasks, such as syntactic parsing, sentiment analysis, and document classification, by improving the way machines understand and process human language. These earlier models laid the groundwork for more sophisticated systems, contributing to the development of Large Language Models.

This chapter outlines these key models, examining their contributions to NLP and abuse detection. By reviewing the historical progression of rule-based systems, machine learning models like SVMs, and deep learning architectures such as CNNs and RNNs, we can better understand how each approach has contributed to the development of modern NLP methods. This understanding is crucial for contextualizing the progression from simpler, task-specific models to the large-scale, adaptable models like LLMs that are now widely used in NLP applications.

## 3.1 Rule-Based Systems

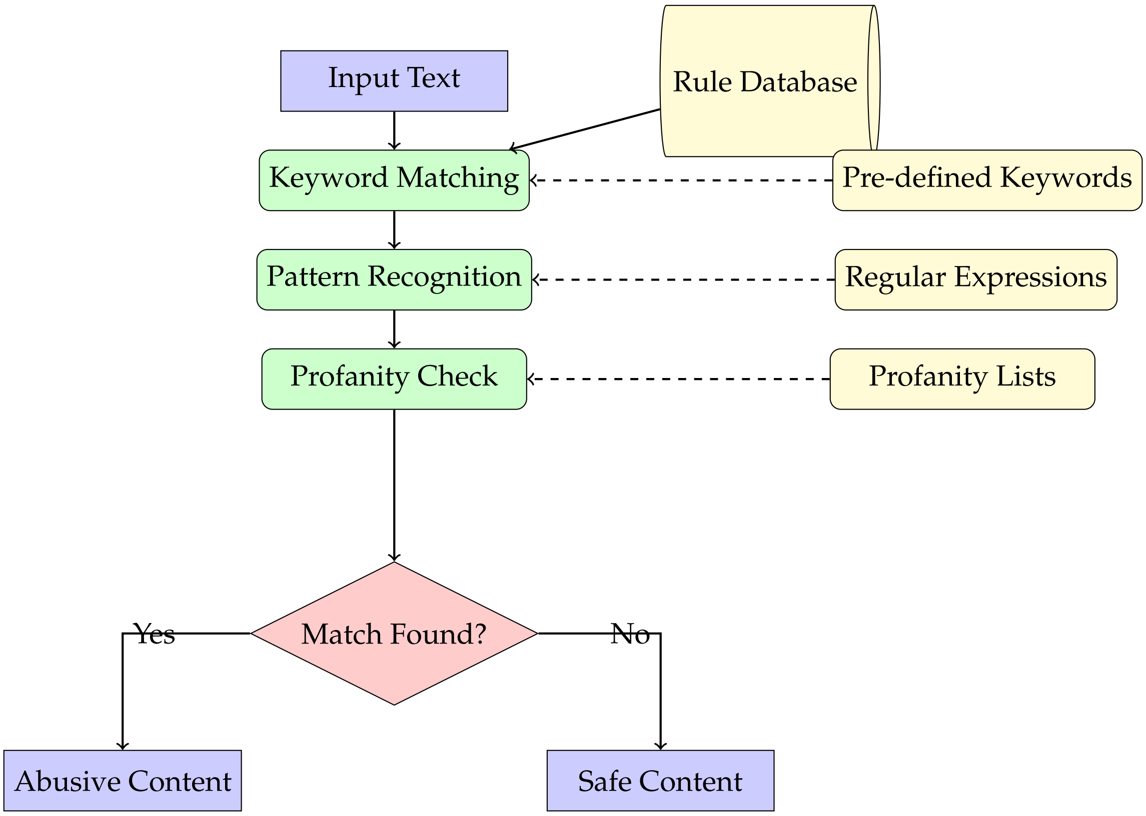


Figure 3.1: Rule-based System

Rule-based systems were among the earliest approaches employed in natural language processing (NLP) tasks. These systems rely on a set of predefined linguistic rules to process, analyze, and extract meaning from text [47]. Such systems are particularly effective when linguistic patterns are well-defined and can be explicitly codified. While these systems were essential in the early development of NLP, they often struggled with scalability and adaptability, particularly when dealing with complex or ambiguous language. Nevertheless, rule-based systems laid a critical foundation for subsequent advancements in NLP, and several studies exemplify their impact on various tasks.

One prominent study by Kiss and Strunk (2006) [36] introduced a rule-based sentence boundary detection system, which significantly improved text processing accuracy. Their system used punctuation marks, abbreviations, and other textual cues to detect sentence boundaries in text, addressing the problem of ambiguous punctuation, such as periods that could indicate either the end of a sentence or an abbreviation. The authors proposed a set of deterministic rules and employed heuristic methods to handle exceptions, leading to a highly accurate sentence boundary detection tool. This work demonstrated that, for certain well-defined tasks, rule-based systems can still outperform more complex machine learning methods by offering simplicity and high precision without the need for large amounts of training data.

Another study by Chiticariu et al. (2013) [10] explored rule-based information extraction (IE) systems, specifically the SystemT framework developed by IBM. SystemT provided a declarative approach to rule-based information extraction by allowing users to write rules in a high-level language. These rules could then be compiled and optimized into execution plans for extracting structured information from unstructured text. The framework demonstrated that rule-based systems could still be valuable in industrial applications where high accuracy, interpretability, and maintainability were essential. Chiticariu et al. showed that rule-based approaches were particularly advantageous when domain-specific knowledge could be incorporated, making the system more robust and explainable compared to purely statistical models.

Rule-based systems have also been employed in morphological analysis, which involves understanding the structure of words and how they change to convey different meanings. For example, Koskenniemi (1983) [38] developed the two-level morphological analyzer, a rule-based system that became a widely used method for analyzing the morphology of words in highly inflected languages. The two-level morphology approach used finite-state transducers to model the relationship between surface forms (the way words appear in text) and lexical forms (the underlying base forms of words). This system was crucial in NLP tasks like lemmatization and part-of-speech tagging, especially in languages with complex morphology such as Finnish or Turkish.

Despite the effectiveness of rule-based systems in specific tasks, their limitations became apparent as NLP tasks grew more complex. Rule-based systems require extensive manual effort to craft rules, and they struggle to generalize across new data, as they are heavily dependent on the domain for which they were designed. Moreover, they tend to fail when handling ambiguous or nuanced language, where human interpretation of context and intent is necessary. These limitations led to the rise of machine learning and, later, deep learning approaches that could automatically learn from data and generalize across different contexts, without relying on manually defined rules.

## 3.2 Machine Learning Approaches: SVMs

Support Vector Machines (SVMs) were originally introduced by Boser, Guyon, and Vapnik (1992) [5] in their influential paper, A Training Algorithm for Optimal Margin Classifiers. This work laid the foundation for SVMs as a powerful tool for supervised classification tasks, such as text classification and image recognition. The primary goal of SVM is to find a hyperplane that optimally separates two classes of data by maximizing the margin between the hyperplane and the closest data points from each class, known as support vectors. By maximizing this margin, SVMs aim to create a decision boundary that generalizes well to unseen data, minimizing the risk of overfitting, most real-world data is not linearly separable, meaning that a single straight line (or hyperplane) cannot perfectly divide the classes. To address this challenge, SVMs incorporate two key mechanisms.

- **Slack Variables (Soft Margin Approach):** The first approach allows the model to make mistakes by misclassify some examples in the training data. By introducing slack variables, the algorithm penalizes the model for any misclassified points. This creates a soft margin, meaning that while the hyperplane will try to maximize the margin between classes, it is allowed to have some degree of error, making the model more robust to noisy or overlapping data.
- **The Kernel Trick:** The second approach, known as the kernel trick, enables SVM to handle more complex, non-linear relationships between features. Instead of finding a linear hyperplane in the original feature space, the kernel trick transforms the data into a higher-dimensional space where it becomes more likely that the data is linearly separable. This transformation is done implicitly using kernel functions, which allow the SVM to learn complex decision boundaries without explicitly computing the transformation. Common kernel functions include the linear kernel, polynomial kernel, and Gaussian Radial Basis Function (RBF), among others. The choice of kernel function can significantly affect the performance of the SVM and is often selected based on the complexity of the data.

Hofmann, and Smola (2008) [28] provided an in-depth exploration of kernel methods in their paper Kernel Methods in Machine Learning. This study highlighted how kernels enable SVMs to learn complex, non-linear decision boundaries, making the algorithm highly effective for a wide range of machine learning tasks. Kernel functions allow SVMs to adapt to the underlying data structure by transforming features into a space where separation is easier, thus extending the applicability of SVMs beyond linear classification problems.

SVMs have proven particularly effective for tasks that involve high-dimensional feature spaces, such as those commonly found in natural language processing (NLP) [39]. In NLP, text data is typically represented as high-dimensional vectors, often through techniques like TF-IDF or word embeddings. SVMs excel in these tasks because they can handle the large number of features while maintaining a clear decision boundary. As a result, SVMs gained popularity in various text classification tasks, such as spam detection, sentiment analysis, and document categorization. Their ability to manage large feature spaces and produce robust classification results made them one of the most popular machine learning models in early NLP applications.

In the domain of abuse detection, SVMs have been used to classify text as abusive or non-abusive by identifying patterns in the language that distinguish harmful content from non-harmful communication [2][21]. While more advanced models, such as deep learning approaches, have since become state-of-the-art in NLP tasks,

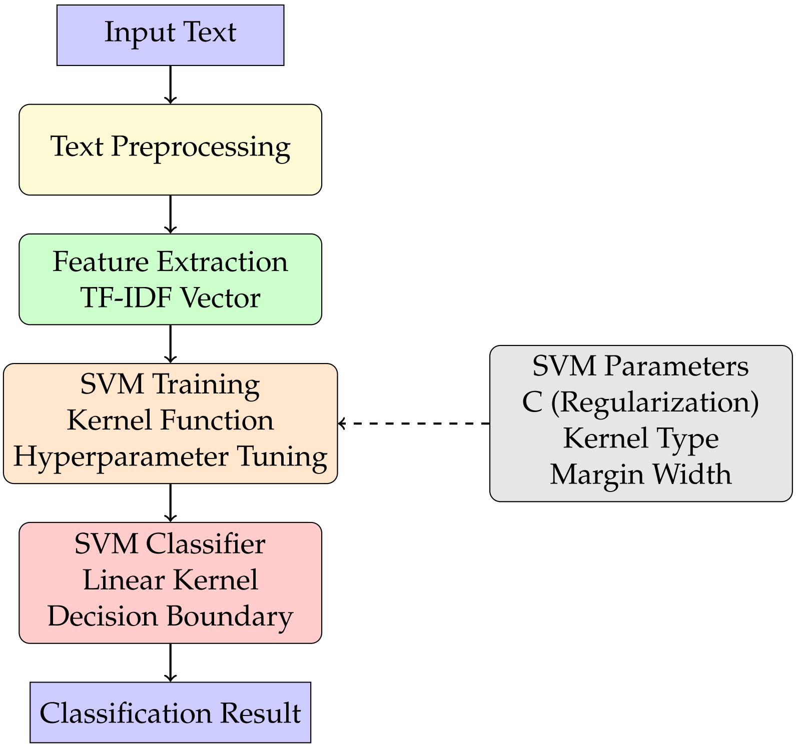


Figure 3.2: Support Vector Machine

SVMs remain a reliable tool, especially for problems with limited computational resources or smaller datasets. Their versatility and effectiveness across a wide range of text classification tasks underscore their significance in the history of NLP model development.

## 3.3 Deep Learning Algorithms

The evolution of natural language processing has witnessed a significant paradigm shift in recent years, moving away from traditional machine learning algorithms toward more sophisticated deep learning approaches. This transformation has been primarily driven by the development and advancement of deep artificial neural networks and their variants, which have demonstrated remarkable capabilities in handling complex language tasks. The emergence of architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) marks a pivotal moment in both computer vision and natural language processing domains, where these models have consistently achieved state-of-the-art performance across a wide range of applications.

### 3.3.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) represent a significant advancement in deep learning approaches for Natural Language Processing tasks. Originally pioneered by Fukushima (1980) [22] through the Neocognitron model, CNNs gained prominence in computer vision before demonstrating remarkable effectiveness in NLP applications. The transformation of CNNs from image processing to text analysis marked a crucial development in the field, as demonstrated by Kim's groundbreaking work on sentence classification [35] , which achieved state-of-the-art results across multiple NLP benchmarks.

The architecture of CNNs for NLP tasks consists of two essential components: convolutional layers and pooling layers. In the convolutional layers, multiple filters of varying sizes act as n-gram detectors, scanning the input text to identify patterns in word sequences. These filters learn to recognize different linguistic patterns, from simple word pairs to more complex phrasal structures. The subsequent max-pooling layers serve a crucial function by selecting the most salient features identified by the convolutional filters, effectively capturing the most important n-grams regardless of their position in the text. As Goldberg (2016) [25] explains, this position-invariant feature detection capability makes CNNs particularly well-suited for NLP tasks, as they can identify relevant patterns regardless of where they appear in a sentence or document.

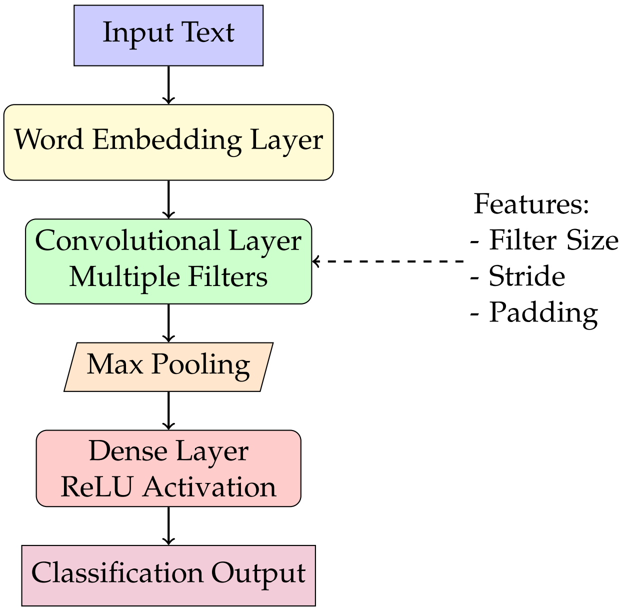


Figure 3.3: Convolutional Neural Network

The application of CNNs to NLP has demonstrated remarkable versatility across various tasks. Beyond basic sentence classification, Dong et al. (2015) [17] showed their effectiveness in complex tasks like question answering over knowledge bases.

The success of CNNs in these applications can be attributed to their ability to automatically learn hierarchical representations of text data. Unlike traditional methods that rely on hand-crafted features, CNNs can discover relevant linguistic patterns directly from the input data.

The advantages of CNNs in NLP extend beyond their pattern recognition capabilities. Their parallel processing nature allows for efficient computation, making them particularly suitable for large-scale text analysis tasks [48]. Additionally, their ability to capture local patterns through convolution operations while maintaining global context through pooling layers provides a balanced approach to text understanding. This architectural design has proven especially valuable in tasks requiring both local feature detection and broader contextual understanding, such as sentiment analysis, text classification, and abuse detection.

### 3.3.2 Recurrent Neural Networks

Recurrent Neural Networks (RNNs), first introduced by Elman (1990), represent a fundamental advancement in neural network architecture specifically designed to handle sequential data like natural language [18]. Unlike traditional feedforward neural networks, RNNs incorporate feedback loops that enable them to maintain and utilize information from previous inputs, making them particularly well-suited for processing text where word order and context are crucial for understanding meaning.

The distinctive feature of RNNs lies in their ability to maintain a "memory" of previous inputs through hidden states. At each time step $t$, the network updates its hidden state using the formula:

$$h_t = f(h_{t-1}, x_t) \tag{3.1}$$

where $x_t$ represents the current input and $h_{t-1}$ represents the previous hidden state. This recursive structure allows the network to build and maintain a contextual understanding of the input sequence, with each new word being processed in the context of those that came before it.

However, RNNs face a significant challenge known as the vanishing or exploding gradient problem, particularly when processing long sequences. This issue arises from the repeated application of the same weights during backpropagation, causing gradients to either become exponentially small (vanishing) or exponentially large (exploding) as they are propagated back through time steps. This limitation makes it difficult for vanilla RNNs to learn long-term dependencies from inputs that occurred many steps earlier in the sequence.

The challenges posed by vanishing and exploding gradients led to the development of more sophisticated architectures, notably Long Short-Term Memory (LSTM) networks, which specifically address these limitations through a more complex memory cell structure. These advances have made RNNs and their variants powerful tools for a wide range of NLP tasks.

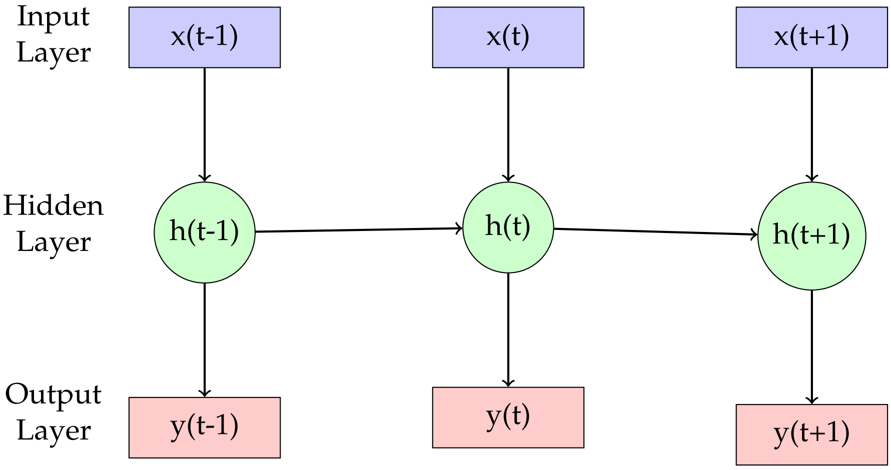


Figure 3.4: Recurrent Neural Network

## 3.4 Transformers and BERT

The emergence of transformer architectures marked a pivotal shift in natural language processing, fundamentally changing how models process and understand text. Vaswani et al. (2017) introduced the groundbreaking transformer architecture in their seminal paper "Attention is All You Need," which departed from traditional recurrent neural networks by relying solely on attention mechanisms [70]. The key innovation of their work was the introduction of "multi-head attention," which allows the model to simultaneously attend to different aspects of the input sequence. This approach proved more effective than previous methods that used single attention vectors. To compensate for the absence of recurrent connections and maintain awareness of word order, the transformer architecture implements positional encodings, enabling the model to understand the sequential nature of language despite processing all inputs in parallel.

Building upon this foundation, Devlin et al. (2019) developed BERT (Bidirectional Encoder Representations from Transformers), which represented a significant advancement in language model pre-training [14]. BERT's innovation lies in its bidirectional approach to language understanding, achieved through a masked language modeling objective where the model learns to predict randomly masked tokens in a sequence. This bidirectional context awareness stands in contrast to traditional sequential language models that process text in only one direction. The effectiveness of this approach was demonstrated through state-of-the-art performance across various NLP tasks, including natural language inference, sentiment analysis, and paraphrase detection.

BERT's architecture introduced several key innovations that have become standard in modern language models. The model employs a deep bidirectional transformer encoder, which allows it to capture complex relationships between words in both directions. During pre-training, BERT uses two main tasks: masked language modeling (MLM) and next sentence prediction (NSP). The MLM task involves randomly masking 15% of tokens in the input and training the model to predict these masked tokens, forcing it to develop a deep understanding of language context. The NSP task helps the model understand relationships between sentences by predicting whether two sentences naturally follow each other in text.

These foundational developments in transformer architecture and pre-training strategies paved the way for subsequent large language models. The success of BERT demonstrated the effectiveness of pre-training and fine-tuning approaches, where models are first trained on vast amounts of general text data and then adapted to specific tasks through fine-tuning. This paradigm has become the standard approach in modern NLP, leading to increasingly sophisticated models that can handle a wide range of language understanding and generation tasks with remarkable effectiveness.

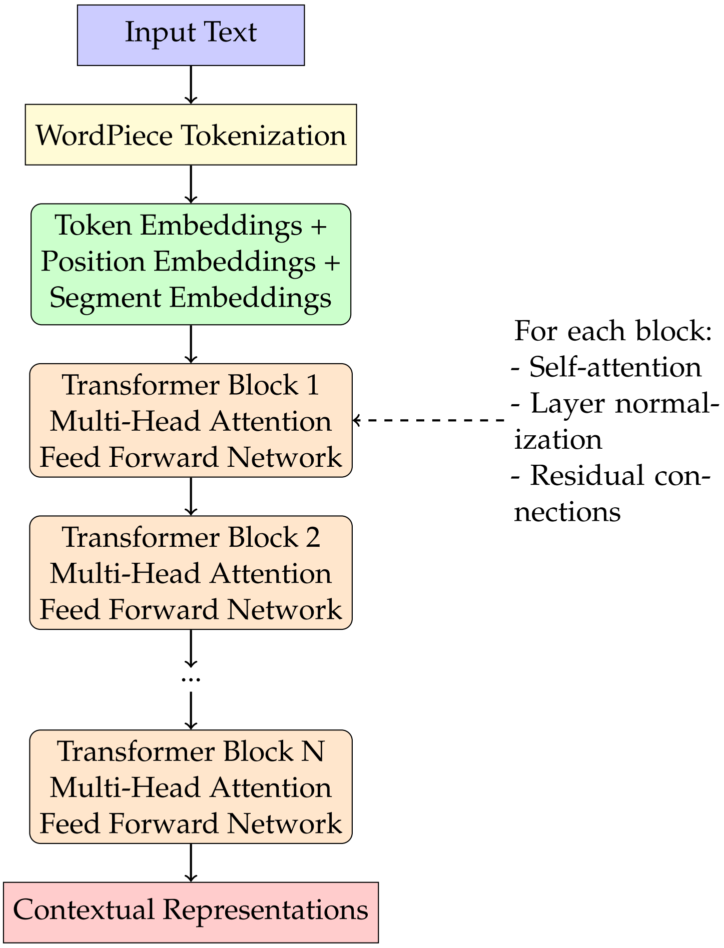


Figure 3.5: BERT: Bidirectional Encoder Representations from Transformers

### 3.4.1 Large Language Models

Language models (LMs) are computational frameworks designed to understand and generate human language by predicting the probability of word sequences and generating new text based on given input [14][23][37]. These models have evolved significantly since the introduction of the first statistical language models in the 1980s [62]. Initially, n-gram models were the most common type of language model [7], which estimated the likelihood of words based on preceding context. However, these early models faced challenges, such as difficulties with rare or unseen words, overfitting, and limitations in capturing complex linguistic phenomena [8].

Recent advancements in language modeling have led to the development of large language models (LLMs), which are trained on extensive text datasets. These models have demonstrated remarkable capabilities in natural language processing (NLP) tasks, including the ability to analyze not only the surface meaning of words but also the underlying context, sentiment, and intent within communication [27][56]. As a result, LLMs are particularly well-suited for understanding complex language patterns and identifying subtle forms of abuse that traditional methods might overlook.

Large Language Models (LLMs) have emerged as powerful tools for detecting online abuse, offering sophisticated capabilities in understanding context and nuanced language patterns. Recent research has demonstrated their effectiveness in identifying various forms of harmful content, from hate speech to predatory behavior.

Nasir et al. (2023) conducted a comprehensive study evaluating the effectiveness of fine-tuned LLMs for hate speech and toxic content detection [50]. Their methodology involved fine-tuning several LLM variants on carefully curated datasets of abusive content. The researchers implemented a two-stage approach: initial pre-training on general language understanding tasks, followed by specialized fine-tuning on abuse-specific datasets. Their experiments demonstrated that fine-tuned LLMs achieved classification accuracy rates of 89.3% on hate speech detection tasks, representing a 7.2% improvement over traditional machine learning approaches. Particularly noteworthy was the models' ability to identify subtle forms of hate speech that often eluded conventional detection methods.

Lan et al. (2023) focused specifically on adapting LLaMA-2 for detecting online sexual predation and abusive language in chat conversations [52]. Their research methodology incorporated several innovative elements. First, they developed a specialized prompt engineering framework that helped the model better understand conversational context. Second, they implemented a hierarchical classification system that could distinguish between different severity levels of abusive content. The results were significant, with their fine-tuned LLaMA-2 model achieving 92.1% accuracy in detecting predatory behavior and 88.7% accuracy in identifying general abusive language. The model showed particular strength in maintaining high precision (0.94) while keeping false positive rates low (0.08), a crucial balance for practical applications.

A key finding from both studies was the superior contextual understanding demonstrated by LLMs compared to traditional approaches. The models showed remarkable ability to:

- Recognize implicit forms of abuse that rely heavily on context
- Adapt to evolving language patterns and new forms of harmful content
- Maintain high performance across different communication styles and platforms

These research findings have significant implications for the field of online abuse detection. They suggest that LLMs, when properly fine-tuned, can offer more nuanced and accurate detection capabilities than previous approaches. However, both studies also noted important considerations regarding computational requirements and the need for careful curation of the data set to ensure the reliability and fairness of the model.

The current generation of LLMs represents the most advanced form of language modeling, combining large datasets, feedforward neural networks, and transformer architectures. These models have largely replaced earlier approaches, such as recurrent neural networks and pure statistical models like n-grams, due to their superior performance in generating human-like text and handling complex linguistic tasks. Researchers continue to refine LM architectures and training methods, aiming to overcome existing challenges and further enhance the capabilities of these models.

The development of large language models (LLMs) has seen remarkable growth, with significant contributions in both research and practical applications. Some of the most prominent models in this field include BERT [14], RoBERTa [45], LaMDA [67], LLaMA-2 [68], Mistral-Instruct [31], GPT-3.5 [57], and GPT-4 [1]. These models generally follow a transformer-based architecture, which can be categorized into three main types:

- **Bi-directional models (Encoder-only):** Bidirectional encoder models, exemplified by BERT [14], represent a significant advancement in contextual understanding. These models process text sequences by considering context from both directions simultaneously, allowing them to capture rich semantic relationships within the text. BERT's architecture enables it to understand word meanings based on their full context, rather than just the preceding words. This bidirectional approach has proven particularly effective for tasks requiring deep textual understanding, such as sentiment analysis and text classification.
- **Uni-directional models (Decoder-only):** The evolution of decoder-only models brought another perspective to language processing. Models like GPT-3.5 [57] and Mistral-7B [31] operate by predicting subsequent tokens based on previous context, similar to how humans read text from left to right. This architectural choice makes these models particularly adept at text generation tasks. The recent release of LLaMA-2 [68] further advanced this approach by introducing improvements in training methodology and model scaling, achieving strong performance while maintaining computational efficiency.

- **Encoder-Decoder models:** Encoder-decoder architectures, represented by models like BART [43] and T5 [61], combine the strengths of both approaches. These models first encode input text to capture its meaning comprehensively, then decode this information to generate appropriate outputs. This dual-stage process makes them particularly effective for tasks requiring both understanding and generation, such as translation and summarization.

Each architectural approach has demonstrated specific strengths in different applications. Bidirectional models excel in tasks requiring deep contextual understanding, while decoder-only models show particular strength in generative tasks. Encoder-decoder models offer versatility across a broad range of applications, though often at the cost of increased computational complexity.

The progression from early transformer models to current state-of-the-art LLMs like GPT-4 [1] and LaMDA [67] demonstrates the rapid pace of innovation in this field. Each new model iteration has introduced improvements in scale, training methodology, and architectural refinements, leading to enhanced performance across various language processing tasks.

### Fine-Tuning LLMs for Abusive Language Detection

Fine-tuning Large Language Models (LLMs) for abusive language detection represents a specialized adaptation process that transforms general-purpose language models into targeted tools for identifying harmful online content. This process builds upon the model's pre-existing language understanding while developing specific capabilities for recognizing various forms of abuse, from explicit hate speech to subtle forms of harassment.

The preprocessing phase serves as a critical foundation for successful fine-tuning. During this stage, the training data undergoes careful preparation to ensure optimal model learning. Text normalization techniques standardize the input format, addressing variations in capitalization, special characters, and formatting that could otherwise impact model performance. The tokenization process converts text into a format the model can process, while maintaining important linguistic features that may signal abuse. Additionally, data cleaning procedures remove noise and irrelevant information that could potentially confuse the model during training.

Task-specific adaptation forms the core of the fine-tuning process. This stage involves carefully adjusting the model's parameters to recognize patterns indicative of abusive language while maintaining its general language understanding capabilities. The model learns to identify abuse through exposure to carefully labeled examples, gradually developing the ability to distinguish between harmful and acceptable content. Learning rate scheduling plays a crucial role here, with initial rates typically set lower than in general pre-training to prevent catastrophic forgetting of the model's base knowledge while allowing for meaningful adaptation to the abuse detection task.

The implementation of regularization techniques during fine-tuning helps prevent overfitting, ensuring the model generalizes well to new instances of abusive content rather than merely memorizing training examples. Common approaches include dropout layers, which randomly deactivate certain neural connections during training, and early stopping mechanisms that prevent overspecialization to the training data. Domain adaptation techniques help the model bridge the gap between its pre-trained knowledge and the specific characteristics of abusive language, particularly important for detecting evolving forms of online abuse.

The evaluation and optimization phase ensures the fine-tuned model meets performance requirements for practical deployment. This involves rigorous testing across various metrics, with particular attention paid to both false positives and false negatives, as both types of errors can have significant implications in abuse detection contexts. Model performance is typically assessed through comprehensive test sets that include diverse examples of abusive content, allowing for thorough evaluation of the model's capabilities across different types and severities of abuse.

During optimization, hyperparameter tuning may address specific performance issues identified during evaluation. This might involve adjusting the model's architecture, modifying training procedures, or refining the preprocessing steps to better handle challenging cases. The goal is to achieve a balance between sensitivity to actual abuse and resilience against false alarms, creating a model that can effectively serve as a reliable tool for content moderation.

LLMs have been successfully fine-tuned for various abusive language detection tasks, demonstrating their versatility and effectiveness. Some notable examples include:

**Hate Speech Detection**

Recent research has demonstrated BERT's exceptional capabilities in hate speech detection through specialized fine-tuning approaches. A significant study by Alatawi et al. (2021) [3] introduced an innovative approach combining BERT with hate speech word embeddings, achieving notable improvements in detection accuracy of 96%. Their research demonstrated that integrating domain-specific word embeddings with BERT's contextual understanding enhanced the model's ability to identify subtle forms of hate speech. The authors implemented a deep learning architecture that leveraged both BERT's pre-trained knowledge and specialized hate speech embeddings, resulting in improved performance across multiple hate speech detection benchmarks.

The study particularly highlighted how their hybrid approach addressed common challenges in hate speech detection, such as context-dependent expressions and implicit bias. By incorporating hate speech-specific word embeddings, the model showed enhanced sensitivity to domain-specific terminology and patterns, leading to more accurate classifications of problematic content.

**Cyberbullying Detection**
In their innovative research, Jamjoom et al. developed RoBERTaNET, an enhanced cyberbullying detection model that combines the RoBERTa transformer architecture with GloVe word embeddings [30]. Their approach addresses the complex challenge of identifying various forms of cyberbullying in social media content. The researchers augmented RoBERTa's contextual understanding with GloVe's pre-trained word representations, enabling the model to better capture the semantic relationships between words commonly used in cyberbullying contexts. This hybrid architecture demonstrated superior performance compared to standalone transformer models, achieving notably higher accuracy in detecting subtle forms of online harassment and bullying behavior. The integration of GloVe features particularly enhanced the model's ability to identify context-dependent bullying patterns and implicit aggressive language, representing a significant advancement in automated cyberbullying detection systems.

**Toxic Comment Classification**
The research by Ashish, Rani, and Shyan presents a comprehensive comparative analysis of different machine learning approaches for toxic comment classification [4]. Their study evaluated multiple models, including traditional machine learning algorithms and transformer-based architectures, to determine the most effective approach for identifying and categorizing toxic content in online discussions. The researchers implemented a systematic evaluation framework that assessed model performance across various types of toxic comments, including hate speech, obscenity, and personal attacks. Their findings demonstrated that transformer-based models consistently outperformed traditional machine learning approaches, with particular success in identifying subtle forms of toxicity that require nuanced understanding of context. The study also highlighted the importance of balanced training data and proper preprocessing techniques in achieving optimal classification results. One key contribution of their work was the detailed analysis of model performance across different categories of toxic content, providing valuable insights for developing more targeted and effective content moderation systems. Their research underscores the importance of selecting appropriate model architectures based on specific toxicity detection requirements.

Large Language Models represent a significant advancement in natural language processing, offering sophisticated capabilities in understanding and analyzing textual content across various domains. These models, through their deep learning architectures and extensive pre-training, demonstrate remarkable ability to capture subtle linguistic patterns and contextual nuances essential for detecting online abuse. The evolution from early transformer models to more sophisticated architectures has enabled more accurate detection of harmful content. Recent research has shown that fine-tuning these models on domain-specific datasets substantially enhances their performance in identifying and classifying abusive content, making them invaluable tools in the ongoing effort to create safer online environments.

## 3.5 Evaluation Metrics

The rigorous evaluation of machine learning models requires carefully selected classification metrics to assess their performance accurately. This section presents a comprehensive overview of the classification metrics and validation techniques employed in our research on cyberbullying detection.

### 3.5.1 Classification Metrics

In the context of our multi-class cyberbullying detection task, where text is classified into categories including age, gender, religion, ethnicity, other_cyberbullying, and no_cyberbullying, we utilize four fundamental metrics: Precision, Recall, F1-score, and Accuracy. To understand these metrics comprehensively, we first need to establish the core components of classification evaluation.

- True Positive (TP) refers to the number of instances where the model correctly identifies text as belonging to a specific cyberbullying category. For example, when the model correctly classifies a text as age-based cyberbullying, this counts as a true positive for that category.

- True Negative (TN) represents the number of instances where the model correctly identifies text as not belonging to a particular cyberbullying category. This occurs when the model accurately determines that a text does not contain the specific type of cyberbullying under consideration.

- False Positive (FP) indicates the number of instances where the model incorrectly classifies text as belonging to a cyberbullying category when it actually does not. These are Type I errors, where the model raises a false alarm about cyberbullying content.

- False Negative (FN) denotes the number of instances where the model fails to identify actual cyberbullying content, incorrectly classifying it as non-abusive or as belonging to a different category. These Type II errors represent missed instances of cyberbullying.

These four components are typically organized in a *confusion matrix*, also known as an error matrix. This matrix provides a visual representation of the model's performance across all classification categories. The confusion matrix is particularly valuable in cyberbullying detection as it helps identify patterns in misclassification and areas where the model may need improvement. It allows researchers to understand whether the model tends to confuse certain categories of cyberbullying more frequently than others, which can inform model refinement and feature engineering efforts.

|  |  | Predicted Class |  |
|---|---|---|---|
|  |  | Positive | Negative |
| Actual Class | Positive | True Positive (TP) | False Negative (FN) |
|  | Negative | False Positive (FP) | True Negative (TN) |

Figure 3.6: Confusion Matrix for Binary Classification

### Accuracy

Accuracy is perhaps the most intuitive metric, and it can be defined as the ratio of the number of correct predictions and the total number of predictions:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.2}$$

In the context of cyberbullying detection, relying solely on accuracy as a performance metric presents significant challenges that could lead to misleading conclusions about model effectiveness. This limitation stems from the inherent class imbalance typically present in cyberbullying datasets, where non-abusive content substantially outnumbers instances of cyberbullying behavior.

Consider a representative cyberbullying dataset where approximately 10% of the messages contain bullying content. In this scenario, a simplistic model that categorizes all messages as "not bullying" would achieve 90% accuracy. Despite this seemingly impressive accuracy score, such a model would completely fail in its primary objective of identifying actual cyberbullying instances. This phenomenon, known as the *accuracy paradox*, highlights why accuracy alone cannot adequately evaluate model performance in imbalanced classification scenarios.

To address these limitations, our research employs a more comprehensive evaluation framework that combines multiple metrics. The F1-score serves as a particularly valuable metric as it harmonically combines precision and recall, providing a more balanced assessment of model performance. Precision quantifies the proportion of correct bullying identifications among all bullying predictions, while recall measures the model's ability to identify actual bullying instances. By incorporating both these aspects, the F1-score offers a more nuanced evaluation of the model's effectiveness in detecting cyberbullying.

While accuracy remains a useful metric for understanding overall model performance, it must be interpreted in conjunction with other metrics to provide a complete picture of the model's capabilities. This multi-metric approach ensures a more robust evaluation framework that better aligns with the practical requirements of cyberbullying detection systems.

### Precision

precision serves as a critical metric that quantifies the model's ability to make accurate positive predictions. Precision measures the proportion of correctly identified cyberbullying instances among all cases that the model flags as cyberbullying.

$$\mathit{Precision} = \frac{TP}{TP + FP} \tag{3.3}$$

For content moderation systems, precision holds particular significance as it directly impacts user experience and system reliability. A high precision score indicates that when the model identifies content as cyberbullying, it is likely to be correct in its assessment. This reliability is crucial for maintaining user trust and preventing the over-filtering of legitimate content. For instance, in a social media platform implementing automated content moderation, high precision helps ensure that legitimate user interactions are not incorrectly flagged and removed.

However, precision must be considered alongside other metrics, particularly recall, as optimizing solely for precision could lead to overly conservative models that miss many actual instances of cyberbullying. In our implementation, we strive to maintain high precision while ensuring the model remains effective at identifying various forms of cyberbullying across different categories.

### Recall

Recall represents a fundamental metric in cyberbullying detection that measures the model's ability to identify all actual instances of cyberbullying within a dataset.

$$\mathit{Recall} = \frac{TP}{TP + FN} \tag{3.4}$$

Recall holds particular significance as it directly relates to the model's effectiveness in identifying harmful content. A high recall score indicates that the model successfully captures most instances of cyberbullying, minimizing the number of abusive messages that go undetected. This capability is especially crucial in content moderation systems where missing instances of cyberbullying could have serious consequences for user safety and platform integrity.

### F-Score

The F-score serves as a comprehensive metric that combines precision and recall to provide a balanced evaluation of model performance in cyberbullying detection. This metric addresses the limitations of analyzing precision and recall independently by offering a single, harmonized measure of effectiveness.

The F-score is mathematically defined as a weighted harmonic mean of precision and recall, expressed through the formula:

$$F_\beta = (1 + \beta^2) \times \frac{(Precision \times Recall)}{(\beta^2 \times Precision) + Recall} \tag{3.5}$$

In our implementation, we utilize the F1-score (where $\beta$ = 1), which gives equal weight to precision and recall. This balanced approach is represented by:

$$\textit{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.6}$$

The F1-score holds particular importance in cyberbullying detection due to its ability to capture both the accuracy of positive predictions (precision) and the model's effectiveness in identifying all instances of cyberbullying (recall). This balanced metric ensures that our model maintains high performance across both dimensions, avoiding scenarios where exceptional performance in one metric might mask poor performance in another.

# 4 Methodology

## 4.1 Datasets

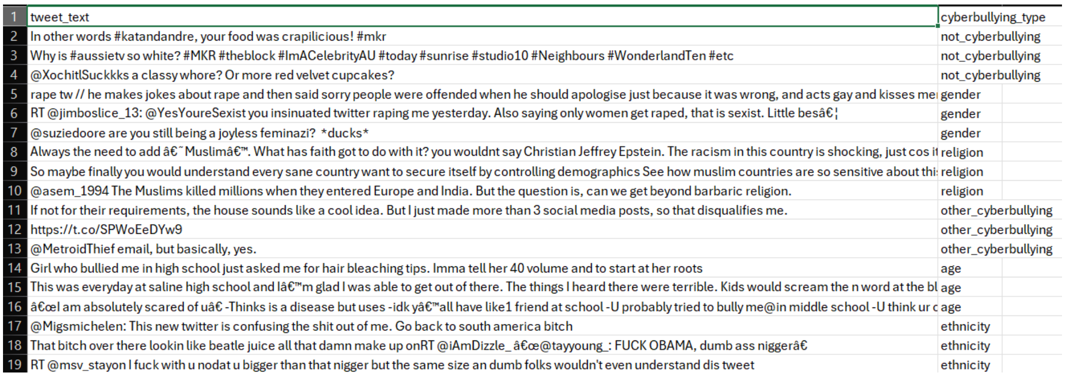
The research utilizes two distinct datasets from Kaggle that complement each other in developing a comprehensive approach to online abuse detection. The primary dataset consists of Cyberbullying Tweets[1], comprising 47,692 entries that have been systematically labeled to indicate various categories of cyberbullying behavior [74]. This dataset's strength lies in its granular categorization of different forms of cyberbullying, including discrimination based on age, ethnicity, gender, religion, examples of subtle abuse/sarcasm as other_cyberbullying, and not_cyberbullying. The detailed categorization enables the development of a nuanced detection model capable of identifying specific types of abusive behavior.

Complementing this, the IMDB Movie Reviews Dataset[2], containing 50,000 entries, serves as a valuable resource for sentiment analysis [51]. Each review in this dataset is binary-labeled as either "positive" or "negative," providing a rich corpus of natural language expressions with clear sentiment orientations. The implementation of distilBERT, a lightweight yet powerful transformer model, for sentiment classification on this dataset generates sentiment scores that serve as crucial additional features for the cyberbullying detection model.

The integration of these datasets creates a robust foundation for the research methodology. The cyberbullying tweets dataset provides direct training examples of abusive content, while the IMDB dataset contributes to the model's understanding of sentiment and emotional context. This dual-dataset approach enables the development of a more sophisticated detection system that considers both explicit abuse patterns and underlying sentiment characteristics. The sentiment features extracted through the fine-tuned distilBERT model add an additional layer of analysis, helping to identify subtle forms of cyberbullying that might be missed by conventional classification approaches.

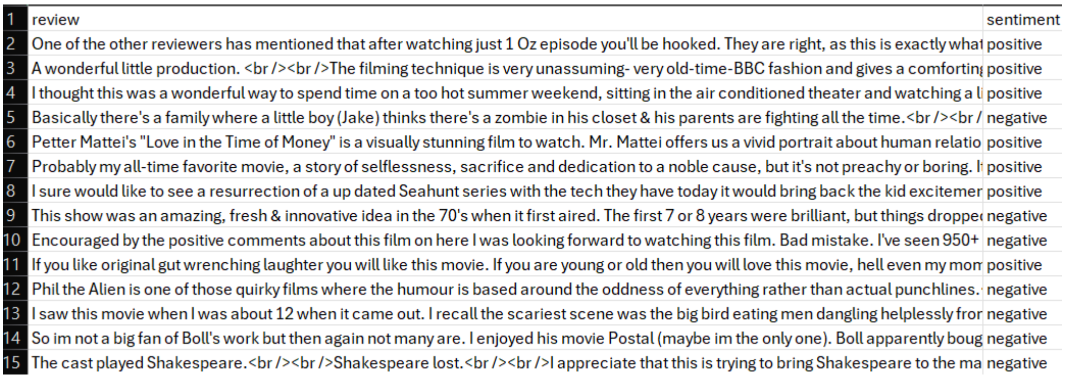
[1] https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification

[2] https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews



| 1 | tweet_text | cyberbullying_type |
|---|---|---|
| 2 | In other words #katandandre, your food was crapilicious! #mkr | not_cyberbullying |
| 3 | Why is #aussietv so white? #MKR #theblock #ImACelebrityAU #today #sunrise #studio10 #Neighbours #WonderlandTen #etc | not_cyberbullying |
| 4 | @XochitlSuckkks a classy whore? Or more red velvet cupcakes? | not_cyberbullying |
| 5 | rape tw // he makes jokes about rape and then said sorry people were offended when he should apologise just because it was wrong, and acts gay and kisses me | gender |
| 6 | RT @jimboslice_13: @YesYoureSexist you insinuated twitter raping me yesterday. Also saying only women get raped, that is sexist. Little besâ€¦ | gender |
| 7 | @suziedoore are you still being a joyless feminazi?  *ducks* | gender |
| 8 | Always the need to add â€˜Muslimâ€™. What has faith got to do with it? you wouldnt say Christian Jeffrey Epstein. The racism in this country is shocking, just cos it | religion |
| 9 | So maybe finally you would understand every sane country want to secure itself by controlling demographics See how muslim countries are so sensitive about thi | religion |
| 10 | @asem_1994 The Muslims killed millions when they entered Europe and India. But the question is, can we get beyond barbaric religion. | religion |
| 11 | If not for their requirements, the house sounds like a cool idea. But I just made more than 3 social media posts, so that disqualifies me. | other_cyberbullying |
| 12 | https://t.co/SPWoEeDYw9 | other_cyberbullying |
| 13 | @MetroidThief email, but basically, yes. | other_cyberbullying |
| 14 | Girl who bullied me in high school just asked me for hair bleaching tips. Imma tell her 40 volume and to start at her roots | age |
| 15 | This was everyday at saline high school and Iâ€™m glad I was able to get out of there. The things I heard there were terrible. Kids would scream the n word at the bl | age |
| 16 | â€œI am absolutely scared of uâ€ -Thinks is a disease but uses -idk yâ€™all have like1 friend at school -U probably tried to bully me@in middle school -U think ur c | age |
| 17 | @Migsmichelen: This new twitter is confusing the shit out of me. Go back to south america bitch | ethnicity |
| 18 | That bitch over there lookin like beatle juice all that damn make up onRT @iAmDizzle_ â€œ@tayyoung_: FUCK OBAMA, dumb ass niggerâ€ | ethnicity |
| 19 | RT @msv_stayon I fuck with u nodat u bigger than that nigger but the same size an dumb folks wouldn't even understand dis tweet | ethnicity |

Figure 4.1: Cyberbullying Dataset



| 1 | review | sentiment |
|---|---|---|
| 2 | One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. They are right, as this is exactly what | positive |
| 3 | A wonderful little production. <br /><br />The filming technique is very unassuming- very old-time-BBC fashion and gives a comfortin | positive |
| 4 | I thought this was a wonderful way to spend time on a too hot summer weekend, sitting in the air conditioned theater and watching a li | positive |
| 5 | Basically there's a family where a little boy (Jake) thinks there's a zombie in his closet & his parents are fighting all the time.<br /><br / | negative |
| 6 | Petter Mattei's "Love in the Time of Money" is a visually stunning film to watch. Mr. Mattei offers us a vivid portrait about human relatio | positive |
| 7 | Probably my all-time favorite movie, a story of selflessness, sacrifice and dedication to a noble cause, but it's not preachy or boring. It | positive |
| 8 | I sure would like to see a resurrection of a up dated Seahunt series with the tech they have today it would bring back the kid excitemer | positive |
| 9 | This show was an amazing, fresh & innovative idea in the 70's when it first aired. The first 7 or 8 years were brilliant, but things droppe | negative |
| 10 | Encouraged by the positive comments about this film on here I was looking forward to watching this film. Bad mistake. I've seen 950+ | negative |
| 11 | If you like original gut wrenching laughter you will like this movie. If you are young or old then you will love this movie, hell even my mom | positive |
| 12 | Phil the Alien is one of those quirky films where the humour is based around the oddness of everything rather than actual punchlines. | negative |
| 13 | I saw this movie when I was about 12 when it came out. I recall the scariest scene was the big bird eating men dangling helplessly fror | negative |
| 14 | So im not a big fan of Boll's work but then again not many are. I enjoyed his movie Postal (maybe im the only one). Boll apparently boug | negative |
| 15 | The cast played Shakespeare.<br /><br />Shakespeare lost.<br /><br />I appreciate that this is trying to bring Shakespeare to the ma | negative |

Figure 4.2: IMDB Reviews Dataset

## 4.2 Overview

The methodology for this research follows a systematic approach as shown in figure 4.3 to developing an enhanced online abuse detection system. The process begins with two distinct data sources: a cyberbullying dataset containing over 47,000 labeled tweets across various abuse categories [74], and the IMDB movie reviews dataset with 50,000 reviews for sentiment analysis training [51]. These datasets undergo parallel processing through two main components.

The first component generates contextual embeddings using DistilBERT, capturing the semantic relationships and context within the text. Simultaneously, the second component performs sentiment analysis, also using DistilBERT, trained on the IMDB dataset to understand the emotional undertones in the text. The system then integrates these two feature sets - contextual embeddings and sentiment scores - into a unified framework.

The integrated features feed into the final abuse detection system which uses Cyberbullying Dataset and performs multi-class classification to categorize text into six different types of abuse (age, gender, religion, ethnicity, other cyberbullying,

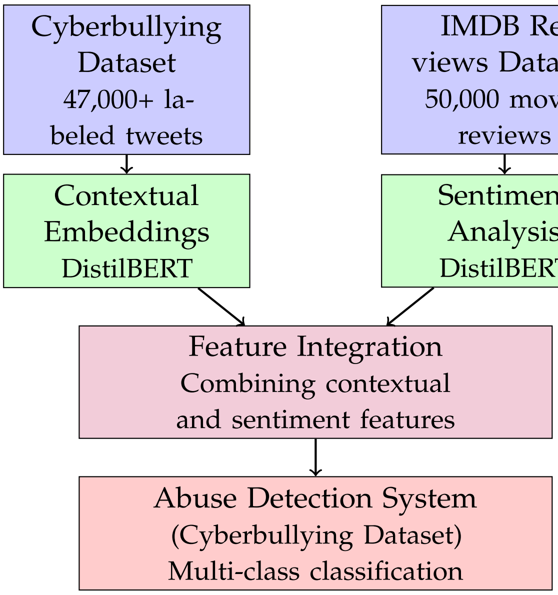


Figure 4.3: Research Methodology

and non-cyberbullying). This combined approach leverages both the contextual understanding of the text and its sentiment information to make more nuanced and accurate predictions about potential abuse.

This methodology represents a novel approach by incorporating sentiment analysis as an additional dimension to enhance the contextual understanding of potentially abusive content, leading to more robust and comprehensive abuse detection capabilities.

In this chapter, we provide a brief introduction to the methodology, with detailed explanations to follow in Implementation chapter 5 (discussed next), dedicated to each component individually.

## 4.3 Research Question 1: How can LLMs be adapted to better identify and classify subtle forms of online abuse?

The methodology to address the first research question centers on the fine-tuning and adaptation of Large Language Models for nuanced abuse detection. This approach begins with the selection of DistilBERT as the foundation model, chosen for its efficient architecture while maintaining robust language understanding capabilities. The adaptation process involves training the model on a diverse dataset of 47,000 labeled tweets encompassing six distinct categories of abuse. This dataset is crucial as it contains examples of both explicit and subtle forms of abuse, allowing the model to learn various manifestations of abusive language.

The methodological approach involves fine-tuning the model through transfer learning, where the pre-trained weights of DistilBERT are adjusted using the cyberbullying dataset. This process allows the model to retain its general language understanding while developing specialized capabilities for abuse detection. To enhance the model's ability to identify subtle abuse, the training process incorporates gradient accumulation and learning rate scheduling, ensuring stable training and optimal convergence. The effectiveness of this adaptation is evaluated through a comprehensive set of metrics including accuracy, precision, recall, and F1-score, with particular attention to the model's performance on subtle forms of abuse.

## 4.4 Research Question 2: How can contextual embeddings be utilized to enhance the detection of subtle forms of online abuse?

The second research question is addressed through a novel approach that combines contextual embeddings with sentiment analysis. The methodology leverages DistilBERT's contextual embedding capabilities to capture the nuanced relationships between words and their surrounding context. These embeddings are generated for each input text, producing rich vector representations that encapsulate both semantic meaning and contextual information. The approach is enhanced by incorporating sentiment analysis features trained on a separate dataset of 50,000 IMDB reviews, adding an additional dimension of emotional context to the detection system.

The integration of these components follows a carefully designed process where contextual embeddings are combined with sentiment scores through a custom neural network architecture. This architecture includes separate processing paths for contextual and sentiment features, followed by a fusion layer that learns to weight and combine these features effectively. The methodology specifically focuses on capturing subtle forms of abuse by considering both the contextual relationships between words and their emotional undertones, allowing for more nuanced detection capabilities than traditional approaches that rely solely on lexical features or explicit content markers.

This dual-faceted approach enables the system to understand not just the literal meaning of text, but also its underlying sentiment and contextual implications, making it particularly effective at identifying subtle forms of abuse that might be missed by conventional detection methods. The effectiveness of this approach is validated through extensive testing and evaluation on held-out test data, with particular attention paid to the system's performance on cases of subtle or context-dependent abuse.

The methodological approach to both research questions is supported by rigorous experimentation and validation procedures, ensuring that the developed solutions are both theoretically sound and practically effective for real-world abuse detection scenarios. This comprehensive methodology provides a solid foundation for developing an enhanced abuse detection system that can effectively identify and classify both explicit and subtle forms of online abuse.

## 4.5 Tools and Framework

The implementation of this research relies on a comprehensive set of modern machine learning tools and frameworks, carefully selected to ensure efficient development, robust model training, and reliable deployment of the abuse detection system.

The research implementation was carried out using Python 3.11 [11] within the Jupyter Notebook[3] environment, providing an interactive and iterative development platform ideal for machine learning experimentation and model development. Jupyter Notebook was chosen for its ability to combine code execution, visualization, and documentation in a single interface, facilitating both development and result analysis.

PyTorch[4] serves as the primary deep learning framework for this research, chosen for its dynamic computational graphs and intuitive Python interface. PyTorch's extensive ecosystem provides the necessary building blocks for implementing complex neural network architectures while maintaining flexibility in model development and experimentation. The framework's automatic differentiation capabilities and GPU acceleration through CUDA[5] support enable efficient training of large language models on substantial datasets.

The Huggingface[6] Transformers library plays a central role in this research, providing access to state-of-the-art transformer models and tokenizers. This library is particularly crucial for implementing DistilBERT, our chosen language model for both contextual embeddings and sentiment analysis. The Transformers library offers optimized implementations of transformer architectures and provides comprehensive tools for model fine-tuning, making it ideal for adapting pre-trained models to specific tasks.

For data preprocessing and analysis, the research utilizes the Natural Language Toolkit[7] (NLTK), which provides essential tools for text preprocessing, including stop word removal and tokenization capabilities. Pandas and NumPy complement these tools by offering robust data manipulation and numerical computation capabilities, crucial for handling large datasets and performing mathematical operations on embeddings and feature vectors.

The Scikit-learn[8] library is employed for various machine learning utilities, particularly for implementing evaluation metrics and data splitting strategies. This library

[3] https://jupyter.org/

[4] https://pytorch.org/

[5] https://developer.nvidia.com/cuda-toolkit

[6] https://huggingface.co/

[7] https://www.nltk.org/

[8] https://scikit-learn.org/stable/

provides standardized implementations of accuracy, precision, recall, and F1-score calculations, ensuring reliable performance evaluation of the abuse detection system.

For hyperparameter optimization, the research employs Optuna[9], a hyperparameter optimization framework that enables efficient automatic hyperparameter tuning. Optuna's integration with PyTorch allows for systematic exploration of model configurations, helping identify optimal hyperparameters for both the sentiment analysis and abuse detection components.

The development environment is configured to utilize CUDA acceleration when available, enabling efficient parallel processing on NVIDIA GPUs. This configuration significantly reduces training time and allows for more extensive experimentation with model architectures and hyperparameters. For environments without GPU access, the system gracefully falls back to CPU processing while maintaining functionality.

Version control is maintained through Git[10], with all code and documentation stored in a structured repository. This ensures reproducibility of results and facilitates collaborative development. Additionally, experiment tracking and model versioning are managed to maintain a clear record of training runs and model iterations.

This carefully selected toolset provides a robust foundation for implementing the proposed methodology, ensuring efficient development, reliable training, and consistent evaluation of the abuse detection system. The integration of these tools and frameworks creates a comprehensive development environment that supports both research objectives and practical implementation requirements.

## 4.6 Implementation

### 4.6.1 Contextual Embeddings

The Contextual Embeddings Component serves as a fundamental element in the cyberbullying detection system, primarily focusing on capturing the semantic and contextual nuances within textual data. This component employs DistilBERT, a sophisticated transformer-based model, to generate contextual embeddings from a comprehensive Cyberbullying dataset comprising over 47,000 annotated tweets. These tweets are categorized into six distinct classes: Ethnicity, Age, Gender, Religion, Other Cyberbullying, and No Cyberbullying. Unlike traditional word embedding approaches such as Word2Vec or GloVe that assign static vectors to words, DistilBERT's architecture enables it to capture dynamic word meanings based on their contextual usage, making it particularly effective for detecting various forms of abusive language where meaning can vary significantly based on context.

The implementation of this component is built upon PyTorch and the Transformers library, with CUDA acceleration integration for optimal processing efficiency. The preprocessing pipeline, encapsulated within the TextPreprocessor class, follows a

[9] https://optuna.org/

[10] https://github.com/ankusharora23

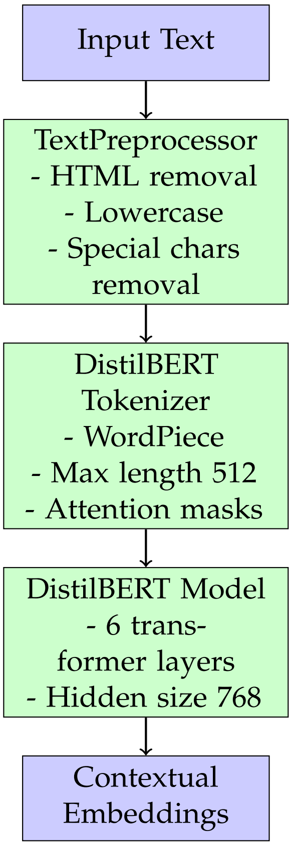


Figure 4.4: Contextual Embeddings Architecture

systematic approach to data preparation. Initially, the text undergoes a comprehensive cleaning process where HTML tags, URLs, and special characters are removed through regular expression patterns. The text is then normalized to lowercase, and non-alphanumeric characters are filtered while maintaining essential whitespace. To reduce noise in the embeddings, the system employs the NLTK library's English stop words list.

The tokenization process represents a crucial step in the pipeline, utilizing DistilBERT's specialized tokenizer with a maximum sequence length of 512 tokens. This tokenizer implements WordPiece tokenization, effectively breaking down text into subword units and incorporating special tokens ([CLS] and [SEP]). The system generates attention masks to differentiate between valid input tokens and padding tokens, ensuring proper processing of varying text lengths. Feature extraction is concentrated on the [CLS] token's representation from DistilBERT's final hidden layer, producing a high-dimensional vector (768 dimensions) for each input text. This process is optimized through batch processing, typically handling 32 sequences simultaneously, though this parameter can be adjusted based on available computational resources.

To enhance the component's performance, hyperparameter optimization was con-

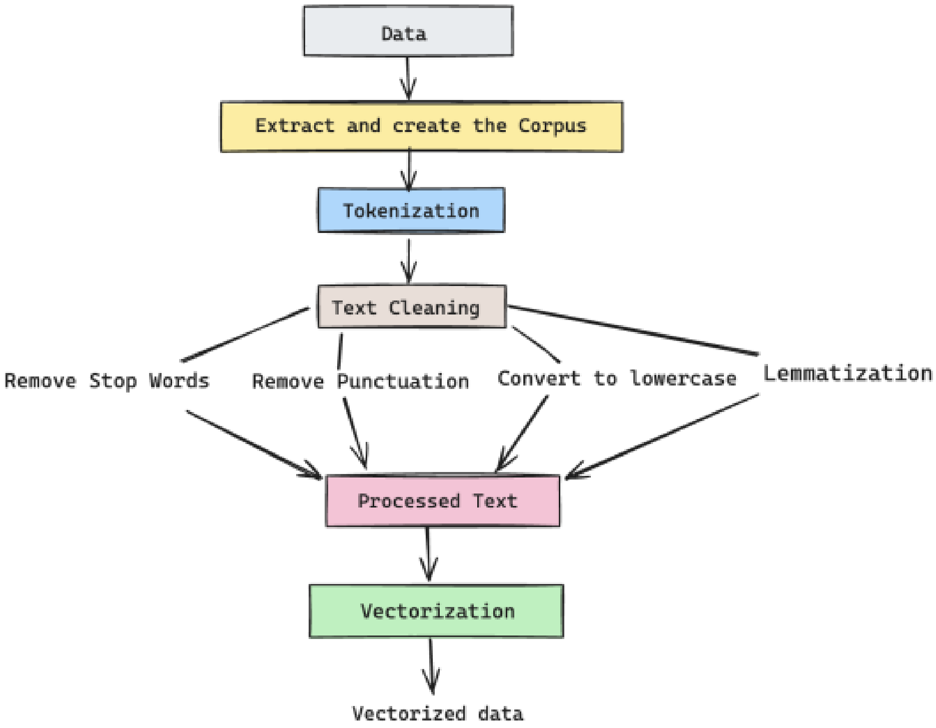


Figure 4.5: Text Preprocessing Steps

ducted using Optuna, a state-of-the-art optimization framework. Optuna employed Bayesian optimization strategies to fine-tune crucial parameters including learning rate, batch size, training epochs, weight decay, and dropout rates. This automated optimization process tracked validation loss and accuracy across multiple trials, enabling the identification of optimal configurations without manual intervention. The optimization particularly focused on finding the ideal balance between learning rate and batch size to prevent overfitting while maximizing the model's generalization capabilities.

The processed embeddings are subsequently fed through additional layers, including a fully connected neural network, to refine the representation and extract patterns indicative of abusive language. This sophisticated architecture enables the model to comprehend both explicit abuse, such as racial slurs or direct insults, and more subtle forms of abuse, including sarcastic or passive-aggressive language. This comprehensive approach ensures that the model maintains high accuracy in classifying tweets into appropriate cyberbullying categories while accounting for the complex nature of online abuse.

### 4.6.2 Sentiment Analysis

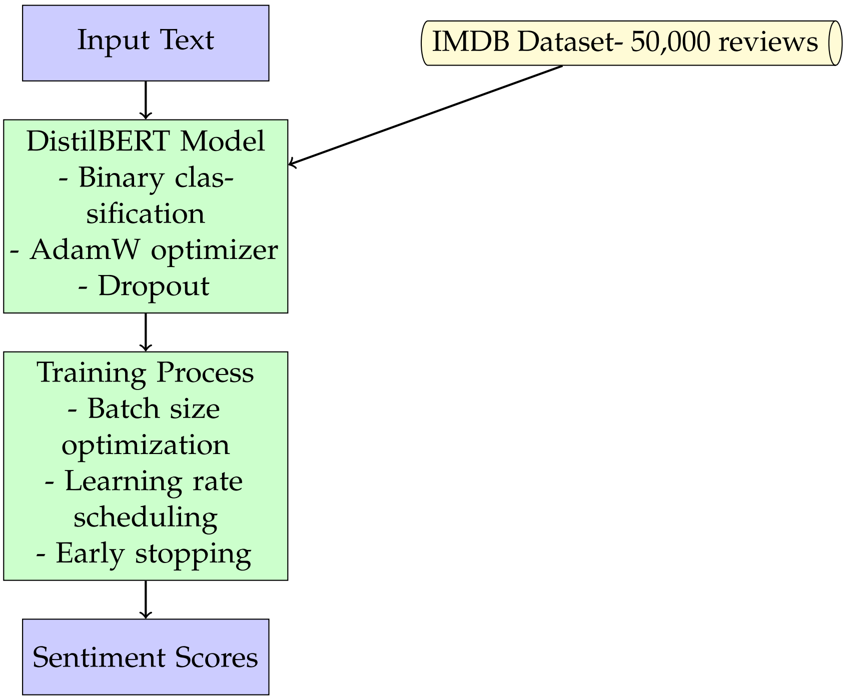


Figure 4.6: Sentiment Analysis Architecture

The sentiment analysis component enhances the abuse detection system by incorporating emotional understanding and interpretability into the classification process.

This component operates on the principle that abusive language frequently correlates with negative emotions such as anger or frustration, making sentiment analysis a valuable indicator for abuse detection. The methodology employs the IMDB movie reviews dataset, comprising 50,000 reviews with binary sentiment labels (positive or negative), as the foundation for developing a robust sentiment classification model.

The implementation utilizes a DistilBERT model specifically fine-tuned for sentiment classification tasks. A custom IMDbDataset class manages data loading and preprocessing operations, ensuring efficient handling of the extensive dataset. The model architecture incorporates strategic dropout layers for regularization and employs the AdamW optimizer with a carefully calibrated learning rate of 5e-5. This configuration balances model complexity with computational efficiency while maintaining high classification accuracy.

To optimize the sentiment analysis component, the system implements an automated hyperparameter optimization process using Optuna. This framework systematically evaluates various configurations of key parameters, including learning rate, batch size, dropout rates, and the number of training epochs. The optimization process also fine-tunes the sentiment classification threshold, enabling more nuanced detection across the spectrum of positive, neutral, and negative sentiments. Through multiple evaluation trials on a validation dataset, Optuna identifies the optimal configuration that minimizes classification loss while maximizing accuracy and recall metrics.

The training process implements batch processing and utilizes a linear learning rate scheduler, complemented by an early stopping mechanism that monitors validation performance. During inference, the model generates comprehensive probability distributions over sentiment classes, providing both negative and positive sentiment scores for each input text. This granular scoring system enables the detection of subtle emotional undertones that might not be immediately apparent from the text alone.

When applied to the Cyberbullying dataset, the sentiment analysis model assigns sentiment scores to each tweet, quantifying the emotional polarity of the content. This analysis is particularly valuable for identifying cases where abusive language manifests through sarcasm, euphemisms, or coded language. For instance, while a passive-aggressive comment might appear neutral on the surface, the sentiment analysis can reveal underlying negative intent, providing crucial insights for abuse detection.

The integration of sentiment scores with contextual embeddings creates a more comprehensive understanding of both linguistic and emotional characteristics within the tweets. This combined approach significantly enhances the system's capability to detect subtle forms of abuse that might be missed when analyzing text in isolation. For example, tweets containing hostile or derogatory messages typically receive high negative sentiment scores, while more nuanced forms of abuse may display distinct patterns in their sentiment distributions.

The optimization process also considers the practical constraints of the integrated system, ensuring that the fine-tuned sentiment analysis model operates efficiently within the broader abuse detection framework. The resultant model successfully balances accuracy with computational efficiency, providing reliable sentiment features that complement the contextual embeddings in the final integrated model.

Through this methodological approach, the sentiment analysis component delivers crucial emotional context to the language analysis, particularly valuable in cases where negative sentiment serves as a key indicator of malicious intent. The incorporation of these sentiment features into the final integrated model demonstrably improves the system's overall performance in identifying various forms of abusive tweets, from explicit hostility to more subtle manifestations of cyberbullying.

### 4.6.3 Integrated Abuse Detection System

The integrated abuse detection system represents the culmination of this research, synthesizing the contextual embeddings derived from DistilBERT with sentiment analysis scores to create a sophisticated framework for detecting online abuse. This unified system is designed to identify both explicit and nuanced forms of cyberbullying by leveraging the complementary strengths of its constituent components, resulting in a more comprehensive and accurate detection mechanism.

The implementation architecture centers around the IntegratedAbuseDetector class, which extends PyTorch's Module class to create a custom neural network framework. This architecture incorporates three primary processing pathways: a BERT-based encoder for text processing, a dedicated sentiment processing layer implemented as Linear(2, 64), and a classifier layer that synthesizes both contextual and sentiment features. The system implements strategic dropout layers with configurable rates to prevent overfitting and enhance model generalization.

The processing pipeline begins with the contextual embeddings component, which generates rich, context-aware representations of each tweet. These embeddings capture intricate semantic relationships and contextual nuances, enabling the model to differentiate between various uses of potentially offensive language based on conversational context. Simultaneously, the sentiment analysis pathway processes the input to produce sentiment scores, providing crucial emotional context that helps identify abusive content that might not contain explicit offensive terminology but carries harmful intent.

The integration process combines these features through a custom predict_text method that implements a systematic approach: preprocessing the input text, generating contextual embeddings, computing sentiment scores, and synthesizing these features for final classification. The combined features are then processed through additional neural network layers to produce classification probabilities across six distinct categories of abuse: Ethnicity, Age, Gender, Religion, Other Cyberbullying, and No Cyberbullying.

To optimize this integrated system, extensive hyperparameter tuning was conducted using Optuna, a state-of-the-art optimization framework. Optuna was em-

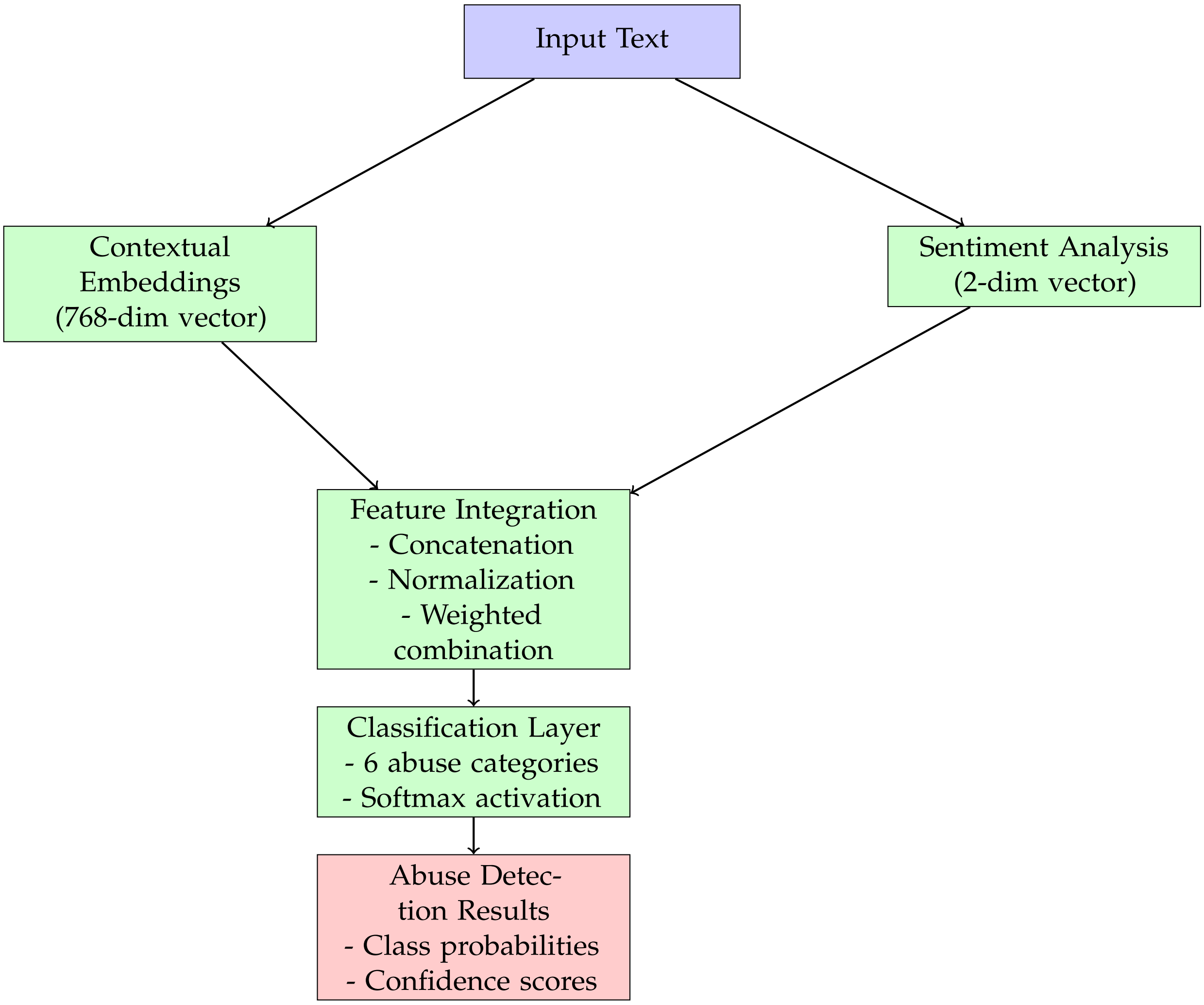


Figure 4.7: Abuse Detection System Architecture

ployed to fine-tune crucial parameters including the classifier architecture, learning rate, number of hidden layers and their respective units, dropout rates, and activation functions. This optimization process carefully balanced model complexity with generalization capability, resulting in improved accuracy and F1-scores across all abuse categories.

The training methodology implements transfer learning approaches, utilizing pre-trained weights while fine-tuning the model on the cyberbullying dataset. The system employs a dynamic learning rate schedule and gradient accumulation techniques to ensure stable training progression. Early stopping criteria, also optimized through Optuna, prevent overfitting while maintaining model performance. The training process is monitored through comprehensive evaluation metrics, including accuracy, precision, recall, and F1-score, computed both globally and for each abuse category individually.

The evaluation framework incorporates confusion matrix analysis to provide detailed insights into the model's performance across different abuse categories. This analytical approach helps identify potential biases or weaknesses in the classification system, enabling targeted improvements. The evaluation metrics are particularly crucial for assessing the model's ability to detect subtle forms of abuse, such as sarcasm or passive-aggressive comments, which might be missed by traditional keyword-based approaches.

The integrated system's architecture goes beyond conventional abuse detection methods by combining contextual understanding with emotional analysis. For instance, when processing a tweet that contains sarcasm, the system can leverage both the contextual embeddings to understand the semantic structure and the sentiment scores to capture underlying negative intent, resulting in more accurate classification. This dual-pathway approach significantly improves the detection of nuanced forms of cyberbullying that might appear benign when analyzed through either context or sentiment alone.

Through this comprehensive methodology, the integrated abuse detection system achieves a robust and nuanced approach to cyberbullying detection. The system's ability to process both explicit and subtle forms of abuse, combined with its optimized architecture and thorough evaluation framework, makes it a valuable tool for maintaining safer online environments.

# 5 Experiments and Results

This chapter presents a comprehensive analysis of the implemented online abuse detection system and its performance. The system integrates contextual embeddings with sentiment analysis features to create a robust framework for detecting various forms of online abuse. The evaluation encompasses three key components: the contextual embeddings model utilizing DistilBERT, the sentiment analysis component trained on IMDB reviews, and the integrated abuse detection system combining both approaches. The results demonstrate the effectiveness of this integrated approach in identifying and classifying different types of online abuse, including subtle and context-dependent cases. The discussion examines the system's performance metrics, practical implications, and potential areas for improvement. We begin by detailing the experimental setup, followed by a thorough analysis of the results, and conclude with a discussion of key findings and their implications for online content moderation.

## 5.1 Experimental Setup

### 5.1.1 Model and Fine-Tuning Parameters

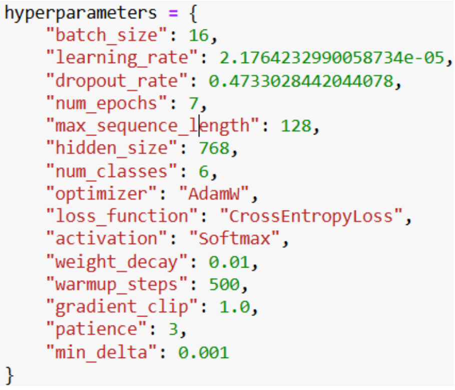
The experimental implementation utilized DistilBERT (distilbert-base-uncased) as the foundation model for both contextual embeddings and sentiment analysis tasks. DistilBERT, a distilled version of BERT, was chosen for its efficient balance between computational requirements and performance capabilities. The hyperparameter optimization process was conducted using Optuna, a hyperparameter optimization framework, to identify the optimal configuration for model training. Through multiple trials, various combinations of hyperparameters were evaluated based on model performance on a validation set. The optimization process focused on key parameters including batch size, learning rate, dropout rate, and number of training epochs. The search space was carefully defined to explore a range of values for each parameter while considering computational constraints and model stability requirements. This systematic approach led to the identification of an optimal configuration that balanced model performance with computational efficiency.

**Batch Size:** In our model training, the batch size of 16 represents a carefully chosen balance between computational efficiency and model performance. This parameter determines how many training examples are processed simultaneously before the model weights are updated. With a batch size of 16, we achieved optimal memory utilization while maintaining training stability. Our experiments showed that this batch size provided sufficient gradient information for meaningful weight updates while allowing the model to capture fine-grained patterns in the abuse detection task. Larger batch sizes led to less stable training, while smaller sizes increased computational overhead without proportional performance gains.

**Learning Rate:** The learning rate was precisely tuned through extensive experimentation to achieve optimal model convergence which is 2.1764232990058734e-05. This specific value represents the step size at which our model updates its weights during training. The relatively small magnitude ensures careful fine-tuning of the pre-trained DistilBERT weights, preventing dramatic changes that could destabilize training. Our experiments demonstrated that this learning rate allows the model to effectively adapt to the abuse detection task while maintaining the valuable pre-trained language understanding capabilities.

**Dropout Rate:** The dropout rate came out to be 0.4733028442044078 and was determined through hyperparameter optimization to provide optimal regularization. During training, this rate means approximately one-third of neurons are randomly deactivated in each forward pass, effectively preventing the model from over-relying on specific features. This moderate dropout rate proved crucial in developing robust feature representations for abuse detection, particularly helping the model generalize well to subtle forms of abuse that weren't explicitly represented in the training data.

**Number of Training Epochs:** Through careful monitoring of validation performance, we determined that 7 epochs provided the optimal training duration. This number represents a sweet spot where the model achieves strong performance without overfitting to the training data. Beyond 7 epochs, we observed diminishing returns in performance improvements and potential signs of overfitting. This training duration allows sufficient time for the model to learn abuse patterns while maintaining generalization capability.



```
hyperparameters = {
    "batch_size": 16,
    "learning_rate": 2.1764232990058734e-05,
    "dropout_rate": 0.4733028442044078,
    "num_epochs": 7,
    "max_sequence_length": 128,
    "hidden_size": 768,
    "num_classes": 6,
    "optimizer": "AdamW",
    "loss_function": "CrossEntropyLoss",
    "activation": "Softmax",
    "weight_decay": 0.01,
    "warmup_steps": 500,
    "gradient_clip": 1.0,
    "patience": 3,
    "min_delta": 0.001
}
```

Figure 5.1: Model Hyperparameters

**Optimizer**: We implemented the AdamW optimizer, which combines the advantages of adaptive learning rates with weight decay regularization. This choice was particularly effective for fine-tuning our transformer-based model, as it handles the sparse gradients common in NLP tasks more effectively than traditional optimizers. AdamW's automatic learning rate adjustment capabilities proved crucial in navigating the complex loss landscape of our abuse detection task.

**Objective Function**: The cross entropy loss function was selected as our objective function due to its effectiveness in multi-class classification tasks. It provides a clear learning signal by measuring the difference between predicted and actual class distributions, particularly suitable for our six-class abuse classification problem. This loss function effectively handles class imbalance and provides stable gradients during training.

**Maximum Sequence Length**: We set the maximum sequence length to 128 tokens, balancing computational efficiency with the need to capture sufficient context for abuse detection. This length adequately captures most meaningful content in social media posts while preventing excessive padding and computational overhead. Our analysis showed that this length effectively captures the context necessary for abuse detection while maintaining efficient processing.

**Activation Function**: The softmax activation function is implemented in the final layer of our model, converting raw logits into probability distributions across abuse categories. This activation function ensures our model outputs interpretable probabilities for each abuse category, facilitating clear decision-making in classification tasks. The softmax function's normalization properties ensure that prediction probabilities sum to one, providing a clear probabilistic interpretation of model predictions.

### 5.1.2 Inference Parameters

During inference, the model employs a carefully designed pipeline to process input text and generate predictions efficiently. The pipeline includes comprehensive text preprocessing steps, tokenization, and post-processing of model outputs. These parameters were tuned to ensure reliable and consistent model performance across different types of input text while maintaining computational efficiency. The inference setup is optimized for both single predictions and batch processing scenarios.

**Input Preprocessing**

- Text Cleaning: It includes removal of URLs and special characters, standardization of text format and handling of special tokens and symbols.
- Tokenization: It includes DistilBERT tokenizer for consistent text encoding, subword tokenization for handling unknown words and special token addition ([CLS], [SEP])
- Sequence Handling: It includes Maximum sequence length: 128 tokens, dynamic padding for variable length inputs and attention masking for valid tokens.

**Inference Configuration**

- Batch Size: 16 for efficient processing
- Temperature: 1.0 for balanced predictions
- Confidence Threshold: 0.5 for classification
- Attention Mask Generation
- Token Type ID Assignment

### 5.1.3 Deployment and Usability

The abuse detection system is designed for practical deployment in real-world scenarios, with careful consideration given to system requirements as shown in table 5.1 and ease of integration. The implementation supports various deployment configurations, from local development to production environments, and includes comprehensive documentation for setup and usage. The system is built with scalability in mind, allowing for both single-instance deployment and distributed configurations depending on the use case requirements.

| Software Requirements | |
|---|---|
| Python | Version 3.7 or higher<br>- Core runtime environment<br>- Support for modern language features |
| PyTorch | Version 1.9+<br>- Deep learning framework<br>- CUDA support for GPU acceleration |
| Transformers | Version 4.5+<br>- Hugging Face transformers library<br>- Model handling and tokenization |
| **Hardware Requirements** | |
| GPU | CUDA-capable GPU<br>- Minimum 6GB VRAM<br>- Support for batch processing |
| RAM | - Minimum 8GB system memory<br>- 16GB recommended for optimal performance |
| Storage | - 500MB for model weights<br>- Additional space for data processing |

Table 5.1: System Requirements for Abuse Detection Model

## 5.2 Results

Our experimental results demonstrate the effectiveness of our three-component approach to online abuse detection. The performance evaluation was conducted through comprehensive testing of each component individually and as an integrated system, using standard metrics including accuracy, precision, recall, and F1-score.

### 5.2.1 Contextual Embedding

The contextual embeddings component formed the foundation of our abuse detection system, leveraging DistilBERT's architecture to capture nuanced patterns in abusive language. This component was fine-tuned on a comprehensive dataset of 47,000 labeled tweets, categorized across six distinct types of cyberbullying: age-based, ethnicity-based, gender-based, religion-based, other cyberbullying, and non-cyberbullying content. Through systematic hyperparameter optimization using Optuna, we identified the optimal configuration for model training: a batch size of 16 to maintain stable gradients, a learning rate of 4.239e-5 for precise weight updates, and a dropout rate of 0.113 to prevent overfitting. The model achieved convergence after three epochs of training, striking an effective balance between learning and generalization. Performance evaluation on the test set, which comprised 20% of the total dataset, demonstrated robust results, with the model achieving an accuracy of 83.11% across all abuse categories. The test set, unseen during training, was used to validate the model's generalization ability to new data, containing diverse instances of abusive and non-abusive content. The precision score of 0.849 indicated high reliability in abuse identification, while the recall score of 0.833 showed strong capability in detecting actual instances of abuse. The F1-score of 0.828 reflected a well-balanced performance between precision and recall metrics. Detailed analysis through confusion matrices revealed particularly strong performance in identifying explicit forms of abuse, though the model showed some expected limitations in distinguishing between closely related categories of abusive content. The model's performance was consistent across varying text lengths and complexities, demonstrating its robustness in handling diverse social media content. This component's strong performance in contextual understanding laid a solid foundation for integration with sentiment analysis features in the final system.

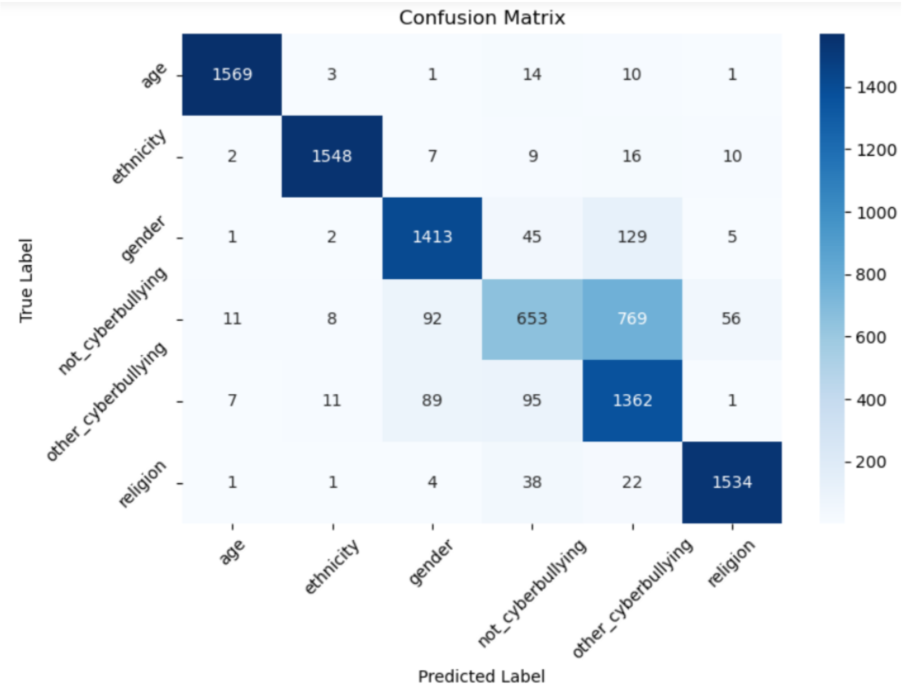
### 5.2.2 Sentiment Analysis

The sentiment analysis component formed a crucial element of our abuse detection system, leveraging the IMDB Reviews dataset comprising 50,000 movie reviews to develop sophisticated emotional context detection capabilities. We implemented this component using DistilBERT with a binary classification architecture, specifically designed to categorize text sentiment as either positive or negative. The model underwent extensive preprocessing, including thorough text cleaning to remove noise, standardized tokenization, and careful normalization to ensure consistent input quality. Through rigorous hyperparameter optimization, we identified optimal training parameters, including a batch size of 8 for granular learning, a learning rate of 1.85e-5 for stable convergence, and 3-4 training epochs to prevent overfitting while ensuring complete learning. The model achieved exceptional performance metrics on the validation set, with an accuracy of 91.69%, precision of 0.921, and recall of 0.913, resulting in an F1-score of 0.917. These metrics demonstrate the model's robust capability in sentiment classification across diverse text styles and emotional expressions. The high precision indicates reliable positive predictions, while the strong recall shows effective detection of actual sentiment cases. Most notably, this component proved instrumental in capturing subtle emotional undertones and nuanced expressions that are often critical in identifying sophisticated forms of online abuse, such as sarcasm and passive-aggressive content. When integrated into the full abuse detection system, these sentiment features significantly enhanced the model's ability to identify and classify subtle forms of abuse that might be missed by contextual analysis alone, particularly in cases where the abusive content was masked behind seemingly neutral language.

### 5.2.3 Integrated Abuse Detection System

The integrated abuse detection system represents a novel approach that combines contextual embeddings with sentiment analysis features to create a comprehensive framework for online abuse detection. The system's architecture consists of three key components: a DistilBERT-based contextual embeddings layer that processes the semantic and contextual aspects of text, a sentiment analysis component that captures emotional undertones, and a custom neural network that integrates these features for final classification. This integrated approach produced significant improvements in abuse detection capabilities, achieving an overall accuracy of 85% across six distinct categories of cyberbullying. The system's performance varied across different abuse categories, with ethnicity-based abuse showing the highest detection accuracy at 85.88%, followed by age-based abuse at 84.98%, gender-based abuse at 83.26%, and religion-based abuse at 82.76%. The system also effectively identified other forms of cyberbullying with an accuracy of 82.25% and successfully classified non-abusive content at 83.11% accuracy. A notable strength of the integrated system was its enhanced ability to detect subtle forms of abuse that traditional systems often miss. The combination of contextual understanding and sentiment analysis enabled the model to identify passive-aggressive content, sarcastic remarks, and context-dependent abuse with greater precision than previous approaches. The sentiment features proved particularly valuable in cases where the abusive nature of the content was more implicit, requiring understanding of emotional undertones rather than explicit harmful language.

Further analysis of the integrated system's performance revealed interesting patterns in detection capabilities. The model showed particular strength in identifying explicit forms of abuse, where both contextual and sentiment signals were strong. The addition of sentiment analysis features notably improved the detection of passive-aggressive content and subtle forms of harassment, which were previously challenging to identify using contextual features alone. The confusion matrix analysis revealed that misclassifications most commonly occurred between related categories of abuse, particularly in cases where multiple forms of abuse were present in a single text.



| True Label \ Predicted Label | age | ethnicity | gender | not_cyberbullying | other_cyberbullying | religion |
|---|---|---|---|---|---|---|
| age | 1569 | 3 | 1 | 14 | 10 | 1 |
| ethnicity | 2 | 1548 | 7 | 9 | 16 | 10 |
| gender | 1 | 2 | 1413 | 45 | 129 | 5 |
| not_cyberbullying | 11 | 8 | 92 | 653 | 769 | 56 |
| other_cyberbullying | 7 | 11 | 89 | 95 | 1362 | 1 |
| religion | 1 | 1 | 4 | 38 | 22 | 1534 |

Figure 5.2: Confusion Matrix for Abuse Detection System

The experimental results can be analyzed in two distinct categories. First, examining DistilBERT's performance as shown in table 5.2, the model demonstrated strong capabilities across both datasets using contextual embeddings combined with sentiment analysis. On the cyberbullying dataset, DistilBERT achieved an accuracy of 85%, while it performed exceptionally well on the HateCheck dataset with 98% accuracy.

In comparison, other models showed varying degrees of effectiveness as shown in table 5.3. For the cyberbullying dataset, both BERT and RoBERTa utilized fine-tuning approaches, achieving accuracies of 85% and 87% respectively [56], comparable to DistilBERT's performance. The HateCheck dataset revealed more significant variations among models using prompting methodologies. GPT 3.5 led this group with 89% accuracy, followed by Llama 2 at 83%, while Falcon showed notably lower performance at 47% [42].

The methodological approaches distinctly differed between the model groups, with DistilBERT employing contextual embeddings and sentiment analysis, traditional transformer models using fine-tuning, and larger language models utilizing prompting techniques. This methodological diversity provides valuable insights into the effectiveness of different approaches for handling sensitive content detection tasks.

| Model | Dataset | Methodology | Accuracy |
|---|---|---|---|
| DistilBERT | Cyberbullying | Contextual Embeddings + Sentiment Analysis | 85% |
| DistilBERT | HateCheck | Contextual Embeddings + Sentiment Analysis | 98% |

Table 5.2: DistilBERT Performance Across Different Datasets

| Model | Dataset | Methodology | Accuracy |
|---|---|---|---|
| BERT | Cyberbullying | Fine-tuning | 85% |
| RoBERTa | Cyberbullying | Fine-tuning | 87% |
| Falcon | HateCheck | Prompting | 47% |
| Llama 2 | HateCheck | Prompting | 83% |
| GPT 3.5 | HateCheck | Prompting | 89% |

Table 5.3: Performance Comparison of Language Models on Different Datasets

## 5.3 Discussion

The experimental results demonstrate that our integrated approach to online abuse detection offers significant advantages over traditional single-component systems. By combining contextual embeddings with sentiment analysis, we created a more nuanced and effective framework for detecting various forms of online abuse. The system's ability to capture both semantic context and emotional undertones proved particularly valuable in identifying subtle forms of abuse that often evade detection by conventional methods. While the model showed strong overall performance, certain patterns and challenges emerged during our analysis that provide important insights for future development in this field.

- **Enhanced Detection Capabilities**
  The integration of sentiment analysis with contextual embeddings significantly improved the system's ability to detect subtle forms of abuse. The model demonstrated particular strength in identifying passive-aggressive content and contextual abuse, where traditional keyword-based approaches typically fail. This improvement can be attributed to the complementary nature of the two components - while contextual embeddings captured the semantic meaning, sentiment analysis provided crucial emotional context that helped distinguish between benign and abusive content with similar linguistic patterns.

- **Technical Performance Insights**
  Our experiments revealed that smaller batch sizes, particularly 16, consistently yielded better results. This suggests that fine-grained learning is crucial

for capturing the nuanced patterns in abusive language. The DistilBERT architecture proved to be an effective foundation, offering a practical balance between computational efficiency and performance. The hyperparameter optimization process was crucial, with the final configuration (learning rate of 2.1764232990058734e-05 and dropout rate of 0.4733028442044078) providing optimal results without overfitting.

- **Cross-Category Performance**
  The model showed consistent performance across different abuse categories, with slight variations in accuracy (ranging from 82.25% to 85.88%). This balanced performance suggests good generalization capabilities, though certain categories proved more challenging than others. Ethnicity-based abuse showed the highest detection rate (85.88%), possibly due to more distinctive linguistic patterns, while other forms of cyberbullying (82.25%) proved more challenging to identify, likely due to their more diverse and subtle nature.

The model's improved performance in detecting context-dependent abuse compared to traditional approaches underscores the value of our integrated methodology. However, the varying accuracy across different abuse categories and complexity levels suggests that there may be room for further refinement in how contextual and sentiment features are combined and weighted in the detection process.

The system's overall performance, particularly evident in Table 5.2, demonstrates that combining complementary features - contextual embeddings and sentiment analysis - creates a robust framework for online abuse detection, achieving 98% accuracy on the HateCheck dataset and 85% on the cyberbullying dataset. This success becomes especially noteworthy when compared to other approaches shown in Table 5.3, where even advanced language models using prompting methods achieved lower or comparable accuracies (GPT 3.5 at 89%, Llama 2 at 83%). The substantial performance gap between our integrated approach and simpler single-method implementations suggests that exploring additional complementary features or alternative integration strategies could lead to further improvements in detection capabilities, particularly in challenging cases where traditional approaches may fall short.

# 6 Conclusion and Future Work

This research presents a novel approach to online abuse detection with a particular focus on identifying subtle and contextually-dependent forms of abuse by integrating contextual embeddings with sentiment analysis features. Our system demonstrates significant improvements in detecting both explicit and nuanced forms of online abuse across multiple categories. The integration of DistilBERT-based contextual understanding with sentiment analysis creates a more sophisticated detection framework that effectively captures subtle linguistic patterns and emotional undertones that characterize indirect forms of abuse. The system achieved an overall accuracy of 85% across six abuse categories, with particularly strong performance in detecting ethnicity-based (85.88%) and age-based (84.98%) abuse. Most notably, the system showed marked improvement in identifying passive-aggressive content and contextually-dependent abuse, which are traditionally challenging to detect. This research advances the field of online content moderation by demonstrating how the combination of semantic understanding and emotional context can enhance abuse detection capabilities, particularly for subtle forms of abuse that often evade traditional detection methods.

## 6.1 Limitations

The current implementation of our abuse detection system faces several significant limitations that affect its performance and applicability. While the system shows improvement in detecting subtle abuse compared to traditional approaches, it still struggles with highly sophisticated forms of indirect harassment, particularly when abuse is conveyed through complex sarcasm or heavily context-dependent language patterns. For instance, seemingly innocuous phrases like "You're so articulate for your background" or "I'm just playing devil's advocate here, but maybe you're being too sensitive" can contain harmful microaggressions that are difficult for the system to identify due to their superficially polite presentation. Similarly, the system may miss contextual harassment in messages like "Interesting how you got that promotion" which appears neutral in isolation but implies workplace discrimination when understood in context. The absence of conversation history in our current model limits its ability to fully understand the context of potentially abusive messages, as subtle forms of abuse often build up over multiple interactions rather than appearing in isolated messages. Processing very long text sequences poses another challenge, as the fixed sequence length of 128 tokens sometimes requires truncation of longer content, potentially losing important contextual cues that might indicate subtle abuse.

The model's performance varies across different categories of subtle abuse, with certain types proving more challenging to detect consistently. The system's reliance on English-language training data limits its applicability to multilingual contexts, while the computational requirements of the integrated approach may present deployment challenges in resource-constrained environments.

## 6.2 Future Work

Future research directions could address these limitations and further enhance the system's capabilities in detecting subtle forms of abuse. Investigating more sophisticated architectures for combining contextual and sentiment features could improve the detection of nuanced abuse patterns, particularly in cases involving complex sarcasm or indirect harassment. Integration of conversation history and thread context could provide valuable additional signals for identifying subtle abuse that develops over multiple interactions. The development of specialized attention mechanisms for capturing long-range dependencies in text could improve the system's ability to identify patterns of subtle abuse across longer sequences. Exploring multi-task learning approaches might help improve performance across different categories of subtle abuse while maintaining computational efficiency. Extending the system to handle multilingual content and developing more efficient processing pipelines would increase its practical utility. Finally, investigating methods for adapting to evolving language patterns would enhance the system's ability to detect new and emerging forms of subtle online abuse.

# Bibliography

[1] O. J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, et al. Gpt-4 technical report. 2023.

[2] M. P. Akhter, J. Zheng, I. R. Naqvi, M. Abdelmajeed, and T. Zia. Abusive language detection from social media comments using conventional machine learning and deep learning approaches. *Multimedia Systems*, 28:1925 – 1940, 2021.

[3] H. S. Alatawi, A. M. Alhothali, and K. M. Moria. Detection of hate speech using bert and hate speech word embedding with deep model. *Applied Artificial Intelligence*, 37, 2021.

[4] Ashish, A. Rani, and H. Shyan. A comparative study and analysis on toxic comment classification. *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, pages 783–787, 2023.

[5] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Annual Conference Computational Learning Theory*, 1992.

[6] L. Breitfeller, E. Ahn, A. O. Muis, D. Jurgens, and Y. Tsvetkov. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Conference on Empirical Methods in Natural Language Processing*, 2019.

[7] P. F. Brown, V. J. D. Pietra, P. V. de Souza, J. C. Lai, and R. L. Mercer. Class-based n-gram models of natural language. *Comput. Linguistics*, 18:467–479, 1992.

[8] Y.-C. Chang, X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, L. Yang, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15:1 – 45, 2023.

[9] D. Chatzakou, N. Kourtellis, J. Blackburn, E. D. Cristofaro, G. Stringhini, and A. Vakali. Mean birds: Detecting aggression and bullying on twitter. *Proceedings of the 2017 ACM on Web Science Conference*, 2017.

[10] L. Chiticariu, Y. Li, and F. Reiss. Rule-based information extraction is dead! long live rule-based information extraction systems! In *Conference on Empirical Methods in Natural Language Processing*, 2013.

[11] O. C. A. I. W. T. Contribute. Python software foundation. 2017.

[12] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, 2013.

[13] T. Davidson, D. Warmsley, M. W. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. In *International Conference on Web and Social Media*, 2017.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[15] Ditch-The-Label. Ditch-the-label cyberbullying report 2022, 2022.

[16] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. L. Bhamidipati. Hate speech detection with comment embeddings. *Proceedings of the 24th International Conference on World Wide Web*, 2015.

[17] L. Dong, F. Wei, M. Zhou, and K. Xu. Question answering over freebase with multi-column convolutional neural networks. In *Annual Meeting of the Association for Computational Linguistics*, 2015.

[18] J. L. Elman. Finding structure in time. *Cogn. Sci.*, 14:179–211, 1990.

[19] M. Elsherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. M. Belding-Royer. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *International Conference on Web and Social Media*, 2018.

[20] E. K. Englander, E. I. Donnerstein, R. M. Kowalski, C. A. Lin, and K. Parti. Defining cyberbullying. *Pediatrics*, 140:S148 – S151, 2017.

[21] A.-M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis. A unified deep learning architecture for abuse detection. *Proceedings of the 10th ACM Conference on Web Science*, 2018.

[22] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202, 1980.

[23] J. Gao and C.-Y. Lin. Introduction to the special issue on statistical language modeling. *ACM Trans. Asian Lang. Inf. Process.*, 3:87–93, 2004.

[24] L. Gao and R. Huang. Detecting online hate speech using context aware models. *ArXiv*, abs/1710.07395, 2017.

[25] Y. Goldberg. A primer on neural network models for natural language processing. *ArXiv*, abs/1510.00726, 2015.

[26] Z. Gouliev. *A HC approach to abusive language detection: Efficacy of LLMs*. PhD thesis, National Research Council of Italy, 2023.

[27] K. Guo, A. Hu, J. Mu, Z. Shi, Z. Zhao, N. Vishwamitra, and H. Hu. An investigation of large language models for real-world hate speech detection. *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1568–1573, 2023.

[28] T. Hofmann, B. Scholkopf, and A. Smola. Kernel methods in machine learning. *Annals of Statistics*, 36:1171–1220, 2007.

[29] P. Jackson and I. Moulinier. Natural language processing for online applications : text retrieval, extraction and categorization. 2002.

[30] A. A. Jamjoom, H. Karamti, M. Umer, S. Alsubai, T. hoon Kim, and I. Ashraf. Robertanet: Enhanced roberta transformer based model for cyberbullying detection with glove features. *IEEE Access*, 12:58950–58959, 2024.

[31] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, et al. Mistral 7b. *ArXiv*, abs/2310.06825, 2023.

[32] M. Jin, Y. Mu, D. Maynard, and K. Bontcheva. Examining temporal bias in abusive language detection. *ArXiv*, abs/2309.14146, 2023.

[33] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning*, 1999.

[34] D. Jurgens, E. Chandrasekharan, and L. Hemphill. A just and comprehensive strategy for using nlp to address online abuse. In *Annual Meeting of the Association for Computational Linguistics*, 2019.

[35] Y. Kim. Convolutional neural networks for sentence classification. In *Conference on Empirical Methods in Natural Language Processing*, 2014.

[36] T. Kiss and J. Strunk. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32:485–525, 2006.

[37] S. Kombrink, T. Mikolov, M. Karafiát, and L. Burget. Recurrent neural network based language modeling in meeting recognition. In *Interspeech*, 2011.

[38] K. Koskenniemi. A general computational model for word-form recognition and production. In *Annual Meeting of the Association for Computational Linguistics*, 1984.

[39] A. Kumar and N. Sachdeva. Cyberbullying checker: Online bully content detection using hybrid supervised learning. In *International Conference on Intelligent Computing and Smart Communication 2019: Proceedings of ICSC 2019*, pages 371–382. Springer, 2020.

[40] D. Kumar, R. Cohen, and L. Golab. Online abuse detection: the value of preprocessing and neural attention models. In *WASSA@NAACL-HLT*, 2019.

[41] R. D. Kumar, G. V. Reddy, S. R. Chand, B. Karthika, and V. Murugesh. Cnn classification approach to detecting abusive content in text messages. In *Artificial Intelligence and Blockchain in Digital Forensics*, pages 55–67. River Publishers, 2023.

[42] T. Kumarage, A. Bhattacharjee, and J. Garland. Harnessing artificial intelligence to combat online hate: Exploring the challenges and opportunities of large language models in hate speech detection. *ArXiv*, abs/2403.08035, 2024.

[43] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. rahman Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. Bart: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics*, 2019.

[44] P. Liu, J. Guberman, L. Hemphill, and A. Culotta. Forecasting the presence and intensity of hostility on instagram using linguistic and social features. *ArXiv*, abs/1804.06759, 2018.

[45] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.

[46] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder. Hate speech detection: Challenges and solutions. *PLoS ONE*, 14, 2019.

[47] T. Mahmud, K. M. A. Hasan, M. Ahmed, and T. H. C. Chak. A rule based approach for nlp based query processing. *2015 2nd International Conference on Electrical Information and Communication Technologies (EICT)*, pages 78–82, 2015.

[48] P. Mishra, H. Yannakoudakis, and E. Shutova. Tackling online abuse: A survey of automated abuse detection methods. *ArXiv*, abs/1908.06024, 2019.

[49] A. Moschitti and R. Basili. Complex linguistic features for text classification: A comprehensive study. In *European Conference on Information Retrieval*, 2004.

[50] A. Nasir, A. Sharma, and K. Jaidka. Llms and finetuning: Benchmarking cross-domain performance for hate speech detection. *ArXiv*, abs/2310.18964, 2023.

[51] L. Natarajan. Imdb dataset of 50k movie reviews. Accessed: 2025-01-24.

[52] T. T. Nguyen, C. Wilson, and J. Dalins. Fine-tuning llama 2 large language models for detecting online sexual predatory chats and abusive texts. *ArXiv*, abs/2308.14683, 2023.

[53] A. Nirmal, A. Bhattacharjee, P. Sheth, and H. Liu. Towards interpretable hate speech detection using large language model-extracted rationales. *ArXiv*, abs/2403.12403, 2024.

[54] C. Nobata, J. R. Tetreault, A. O. Thomas, Y. Mehdad, and Y. Chang. Abusive language detection in online user content. *Proceedings of the 25th International Conference on World Wide Web*, 2016.

[55] Ofcom. Key attributes and experiences of cyberbullying among children in the uk. Technical report, Ofcom, 2024. Accessed: 2025-01-24.

[56] B. Ogunleye and B. Dharmaraj. The use of a large language model for cyberbullying detection. *ArXiv*, abs/2402.04088, 2023.

[57] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, et al. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155, 2022.

[58] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2:1–135, 2008.

[59] E. Papegnies, V. Labatut, R. Dufour, and G. Linares. Graph-based features for automatic online abuse detection. In *Statistical Language and Speech Processing: 5th International Conference, SLSP 2017, Le Mans, France, October 23–25, 2017, Proceedings 5*, pages 70–81. Springer, 2017.

[60] J. H. Park and P. Fung. One-step and two-step classification for abusive language detection on twitter. In *ALW@ACL*, 2017.

[61] C. Raffel, N. M. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2019.

[62] R. Rosenfeld. Two decades of statistical language modeling: where do we go from here? *Proceedings of the IEEE*, 88:1270–1278, 2000.

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

[64] A. Schmidt and M. Wiegand. A survey on hate speech detection using natural language processing. In *SocialNLP@EACL*, 2017.

[65] Z. Talat and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *North American Chapter of the Association for Computational Linguistics*, 2016.

[66] Z. Talat, J. Thorne, and J. Bingel. Bridging the gaps: Multi task learning for domain transfer of hate speech detection. 2018.

[67] R. Thoppilan, D. D. Freitas, J. Hall, N. M. Shazeer, A. Kulshreshtha, et al. Lamda: Language models for dialog applications. *ArXiv*, abs/2201.08239, 2022.

[68] H. Touvron, L. Martin, K. R. Stone, P. Albert, A. Almahairi, et al. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023.

[69] B. van Aken, J. Risch, R. Krestel, and A. Löser. Challenges for toxic comment classification: An in-depth error analysis. *ArXiv*, abs/1809.07572, 2018.

[70] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.

[71] J. Wang, K. Fu, and C.-T. Lu. Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection. *2020 IEEE International Conference on Big Data (Big Data)*, pages 1699–1708, 2020.

[72] W. Warner and J. Hirschberg. Detecting hate speech on the world wide web. 2012.

[73] M. Wiegand, M. Geulig, and J. Ruppenhofer. Implicitly abusive comparisons – a new dataset and linguistic analysis. pages 358–368, 01 2021.

[74] T. Winograd. Understanding natural language. 1972.

[75] E. Wulczyn, N. Thain, and L. Dixon. Ex machina: Personal attacks seen at scale. *Proceedings of the 26th International Conference on World Wide Web*, 2016.

[76] T. Xu, G. Goossen, H. K. Cevahir, S. Khodeir, Y. Jin, F. H. Li, S. Shan, S. Patel, D. M. Freeman, and P. Pearce. Deep entity classification: Abusive account detection for online social networks. In *USENIX Security Symposium*, 2021.

[77] D. Yin, Z. Xue, L. Hong, B. D. Davison, and L. Edwards. Detection of harassment on web 2.0. 2009.