Master's Linguistics Thesis

Context-Aware Hate Speech Detection using BERT: An Investigation with the Contextual Abuse Dataset

Payam Fakhraei

a thesis submitted in partial fulfilment of the requirements for the degree of

MA Linguistics
(Text Mining)

Vrije Universiteit Amsterdam

Computational Lexicology and Terminology Lab
Department of Language and Communication
Faculty of Humanities

Supervised by: Isa Maks
2^{nd} reader: Ilia Markov

Submitted: June 30, 2024

Abstract

The prevalence of toxic language and hate speech on online platforms necessitates the development of automatic detection methods. While most approaches focus on individual comments, this study investigates the impact of incorporating contextual information on the performance of BERT-based models for detecting toxic language in the Contextual Abuse Dataset.

Three experimental conditions are analyzed: using only the comment text, combining the comment with its preceding parent comment, and incorporating the entire conversation thread. The findings reveal that while overall accuracy remains stable, incorporating context improves recall in specific abuse categories. However, models without context demonstrate higher precision, particularly when explicit abuse indicators are present.

This research contributes to understanding the complex interplay between context and hate speech detection models, emphasizing the need for an approach to incorporating relevant contextual information.

Declaration of Authorship

I, -, declare that this thesis, titled *Context-Aware Hate Speech Detection using BERT: An Investigation with the Contextual Abuse Dataset* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a degree degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: 30/4/2024

Signed:

Acknowledgments

List of Figures

Contents

Chapter 1

Introduction

The code repository is available at: https://github.com/pafa3/context-aware-CAD

1.1 The Landscape of Online Hate Speech

Abusive content in the digital landscape is a multifaceted issue. Transcending the boundaries of technology, psychology, sociology, and law, each discipline offers unique perspectives on the challenges and implications of online hate speech. However, despite these perspectives, formulating a universally acceptable definition remains elusive (Fortuna and Nunes, 2018). Research by Jacobs and Potter (1998) and Walker (1994), as elaborated by Davidson et al. (2017), defines hate speech as communication designed to express hatred towards a group or to demean, debase, or insult its members. Some major platforms have responded to this challenge. Prominent social media platforms have been proactive. For example, Facebook defines hate speech as direct aggression towards individuals based on protected aspects such as race, ethnicity, and gender. In contrast, Twitter prohibits the promotion of violence or direct threats based on these protected characteristics (Cortiz and Zubiaga, 2020).

The ripple effects of hate speech are far-reaching, causing both individual harm, such as mental stress, and wider societal damages like inciting violence or public disruption. Marginalized communities are often the main targets of this abuse (Mollas et al., 2022). Highlighting the severity of this situation, the UK government's "Online Harms White Paper" offers startling statistics: around one in five children report instances of cyberbullying, while two-thirds of female journalists have faced online abuse.

1.2 The Role of Context in Addressing Hate Speech

The contextual nature of language is a major challenge when it comes to dealing with online hate speech. Words and phrases, when isolated, can convey drastically different meanings compared to when they are understood within a broader context. A statement that appears benign in isolation might become harmful within a larger conversation (Lemmens et al., 2022). Conversely, what could be deemed offensive might be seen as jest or sarcasm with proper context (Menini et al., 2021).

While the role of context in human communication is widely acknowledged, its significance in automated hate speech detection has been less emphasized. Traditional models are efficient but may stumble when faced with the subtleties introduced by context. This can lead to false positives, flagging innocuous statements, or false negatives, missing genuinely harmful content due to the absence of context (Menini et al., 2021).

The societal and individual impacts of hate speech make it imperative to address this challenge effectively. Thus, this study will focus on utilizing context as a key part of deticting toxic langauge.

1.3 Contextual Abuse Dataset (CAD)

The Contextual Abuse Dataset (CAD) (Vidgen et al., 2021) is a dataset designed to address the complex nature of online conversations, highlighting the importance of context in detecting online abuse. It focuses on English Reddit entries, and each entry is annotated within its conversation thread to better capture the nuances of the conversation.

The labeling process of CAD is thorough and rigorous. Annotators were given instructions to consider the wider conversation when assessing the nature of a statement. The taxonomy of CAD includes categories such as Identity-directed, Person-directed, and Affiliation-directed, providing a detailed view of online abuse. Chapter 3 will discuss the specifics of this dataset further.

Recognizing CAD's potential, my research aims to utilize its contextual content. With the challenges and opportunities presented by the contextual nature of language in mind, this study explores whether context can be used to enhance automated toxic language detection systems.

1.4 Research Questions

The primary objective of this study is to explore the potential benefits of incorporating varying degrees of contextual data into the BERT model, particularly with regard to the detection and categorization of abusive content within the Contextual Abuse Dataset.

1.4.1 Main Research Question

How does the progressive addition of context (from no context to the entire conversation thread) to the BERT model influence its capability to detect and categorize abusive content within the Contextual Abuse Dataset?

1.4.2 Sub-Questions

1. How does the performance of the BERT model vary across the three experimental conditions:

   - When utilizing only the comment (baseline)?
   - When integrating the comment with its preceding parent comment?
   - When taking into account the entire conversation thread?

   This sub-question aims to provide a comparison across the three experimental setups. The objective is to discern the degree to which different layers of contextual information affect the BERT model's precision and efficacy in abuse detection and categorization.

2. With the incremental addition of context to the BERT model (from the baseline to the entire thread), what specific errors are evident in the detection and categorization of abusive content, and what could be the potential causes behind these errors?

Through the identification and examination of these discrepancies, these sub-questions offer insights into the challenges and complexities of context-driven abusive content detection within the BERT paradigm.

In summary, the research questions aim to guide a systematic exploration of the utility and limitations of a context-aware BERT model in detecting and categorizing abusive content. Through these inquiries, the study seeks to contribute empirically grounded insights into the ongoing discourse surrounding the role of context in online hate speech detection and content moderation.

In the next chapter, I will discuss the existing scholarly work on using context while employing hate speech detection model.

Chapter 2

Literature Review

2.1 Introduction

The digital age has enabled the global exchange of ideas and opinions online. Characterized by the proliferation of social media platforms, this epoch has reshaped the paradigms of communication. While this has made it easier to share ideas globally, it has also led to problems such as harmful content, hate speech, and various forms of discrimination. This type of content can negatively impact individuals and society as a whole as they can be aimed at individuals or groups based on attributes such as gender, ethnicity, or religion. And they can have consequential societal and psychological implications (Fortuna and Nunes, 2018; Davidson et al., 2017). Dealing with such a large volume of online interactions manually is impractical, and it creates a need for automated methods to moderate this harmful content. (García-Díaz et al., 2022). This chapter provides an overview of the literature related to the automated detection of abusive content online, especially of those that integrate context, by covering some of its history, methods, and challenges.

2.2 The Evolution of Hate Speech Detection

Online communication platforms have transformed global connectivity, but their rapid growth has also brought significant challenges in content moderation. The vast amount of daily online posts made manual moderation increasingly challenging and emphasized the need for automated solutions (Vidgen et al., 2021).

The initial attempts at automatic content moderation were rudimentary, relying on keyword-based filters to flag or remove potentially offensive content. While these methods could identify potentially offensive terms, they often struggled to differentiate between genuine hate speech and merely offensive language (Davidson et al., 2017; Wulczyn et al., 2017). The presence or absence of specific terms could both help and hinder accurate classification. the subtlety of language, especially when it comes to hate speech, made it clear that the offensiveness of a statement often depended on the context in which it was used. (Baheti et al., 2021).

The rise of machine learning and deep learning brought about a major shift in content moderation. As research progressed, the focus shifted towards understanding the context in which words were used (Agarwal et al., 2022). Models moved away from simple rule-based systems to complex neural networks that could better understand subtleties in language. However, these models faced challenges. The complexities of natural language, especially when attempting to recognize hate speech from mere offensive language, poses significant challenges. The gray areas between these categories made model training and generalization challenging (Baheti et al., 2021).

2.3 The Role of Context

The evolution of online content moderation has underscored the significance of context in understanding and interpreting user-generated content. As online platforms grow rapidly, the diversity and complexity of user interactions necessitate models that can discern not just the explicit meaning of words but also the implicit nuances embedded within the surrounding context. This section explores some recent scholarly papers that have attempted to understand the complexities of context in content moderation. This includes controlling the stance (support, agreement, or disagreement) of generated text to integrating context in various forms into models. By comparing these studies, this review aims to map out the trajectory of research in this domain, explain the methodologies used, and identify potential gaps that can guide future investigations.

In their study, “Just Say No: Analyzing the Stance of Neural Dialogue Generation in Offensive Contexts” Baheti et al. (2021) examines how stance and context intertwine in dialogue models (Baheti et al., 2021). They examine the difficulties in controlling the stance of text generated by AI, especially when it comes to handling offensive language. In particular, they study the challenges inherent in steering the stance of automatically generated text, particularly when working with offensive language.

Baheti et al. (2021) draw a distinction between direct and contextual offensiveness. Direct offensiveness is straightforward, marked by overt derogatory language aimed at specific groups or individuals. In contrast, contextual offensiveness is more subtle. It presents itself when a model’s response, though not explicitly derogatory in isolation, aligns or agrees with a preceding offensive remark.

The study also uncovers a troubling finding: both human participants and dialogue models are much more likely to agree with offensive comments than neutral ones (Baheti et al., 2021). The authors suggest that this trend might be linked to the echo chambers common on social media platforms. In these chambers, users, shielded from dissenting views, often encounter and resonate with opinions mirroring their own, even if they border on offensiveness. When dialogue models are trained on such data, they unintentionally learn and potentially amplify these biases.

To reduce the risk of generating responses that are inappropriate or offensive in specific contexts, the authors introduce the concept of Controllable Text Generation (CTG). Their experiments, particularly with Domain Adaptive Pretraining (DAPT), show some promise, while they also highlight the limitations of such methods. Despite employing advanced CTG techniques, models occasionally falter, underscoring the complexity of using context in dialogue generation.

Baheti et al. (2021) make a notable contribution with the introduction of the TOXICHAT dataset. Unlike other datasets that might be limited to single-turn conversations, TOXICHAT focuses on the complexities of multi-turn Reddit dialogues. In TOXICHAT, context is built by analyzing conversation threads, where the potential offensiveness of each comment is evaluated based on the messages that come before it. This method highlights the importance of dialogue history in determining the stance and offensiveness of responses. A model’s response is analyzed within the conversation thread, showing how previous comments set the stage for later interactions. This approach to context shows the challenges in generating offensive language with neural dialogue systems. It points to the critical need for models that can navigate the nuances of conversation.

2.4 Contextual Understanding in Online Conversations: A Graph-Based Approach

Building on the exploration of stance and offensiveness in generated text, it is essential to consider the broader scope of online conversations. Participatory platforms, characterized by multi-threaded discussions, introduce distinct challenges. In "A Graph-Based Context-Aware Model to Understand Online Conversations" by Agarwal et al. (2022), the authors emphasize the importance of context in interpreting online dialogues.

Agarwal et al. (2022) define context within online dialogues through a graph-based methodology, where context is captured by graph traversals across conversation threads. Their model, GraphNLI, uses these traversals to incorporate broader conversational context by sampling adjacent comments within threads. Their approach to context is by enriching the initial comment's representation with a wider perspective of the dialogues, which resulted in an improved performance in natural language processing tasks such as polarity prediction and hate speech detection compared to their baseline. This approach underlines the importance of a comprehensive view of dialogues for understanding the dynamics of online conversations. They highlight that classifying online comments and replies often depend on an external context that goes beyond immediate preceding or succeeding statements.

Their observation mainly underscores the need for models that can integrate this broader context to improve their performance. As a result, the authors propose GraphNLI, which is a graph-based deep learning architecture. Unlike traditional models that might focus solely on immediate inputs, GraphNLI uses graph to incorporates the broader context of a conversation. Starting from a specific comment and sampling adjacent comments within the same or related threads, the model combines additional embeddings with the initial comment's embedding, which then serves as the basis for various NLP prediction tasks.

The authors' approach to modeling online discussions as directed trees. The original post serve as the root and each subsequent reply creates a directed edge, indicating the flow and structure of the conversation. This representation captures the interconnected nature of online debates, where every comment, except the original post, is a response to another. The model utilizes graph walk techniques to utilize neighboring context in a principled fashion, sampling additional nodes within the global context of online discussions

The model's effectiveness was assessed on two primary tasks: polarity prediction and misogynistic hate speech detection. Polarity prediction, determining whether a reply supports or counters the comment it responds to, is crucial in online debates. Accurate polarity inference can help measure various properties of a debate and can be used in applying argumentation theory techniques to determine which arguments are contested and which are not Saquete et al. (2020). Meanwhile, the challenge of hate speech detection, given its complexities and the diverse ways hate can be expressed online, showcases the model's capabilities. GraphNLI consistently outperformed relevant baselines in both tasks, indicating the potential of context-aware models in understanding online conversations.

2.5 Challenges of using Context in Toxic Language Detection

In their study on preemptive toxic language detection, Karan and Šnajder (2019) explores the intricacies of online conversations, by focusing on Wikipedia comments. They propose that the context for detecting potential toxicity extends beyond individual comments to encompass the entire conversation thread. A "conversation thread" is defined as the sequence of comments and replies that follow an initial post or comment, forming a cohesive discussion. This thread-level information captures the progression and interaction of ideas, which can reveal the evolving context that influences the potential for comments to become toxic. Their research suggests that analyzing the full thread—considering all preceding comments and the initial post—provides a more comprehensive understanding of the conversation's dynamics than examining comments in isolation. This methodological shift towards context aims to improve the accuracy of toxic language detection by considering all preceding comments in the conversation thread.

While the paper advocates for the inclusion of thread-level information, their findings present a different picture. Their experiments reveal that context-sensitive models, despite their promise, did not significantly outperform their context-agnostic counterparts. This finding is particularly intriguing, suggesting that while thread-level context holds potential, its effective integration into predictive models remains an area of study.

2.6 Contextual Nuances in Hate Speech Detection

The intricate relationship between context and hate speech detection is the focal point of the study by Markov and Daelemans (2022). Their research, titled "The Role of Context in Detecting the Target of Hate Speech," explores the challenges and potential solutions of incorporating relevant conversational context into hate speech detection models.

The study emphasizes the deeply contextual nature of hate speech. A statement that appears neutral when examined in isolation may acquire a hateful connotation when considered within a specific context. For instance, the comment "go back home" might be innocuous in many scenarios, but when posted under a news article about refugees, its hateful intent becomes evident. This example underscores the importance of understanding the broader conversational context in which a comment is made.

While previous research has often relied on surface-level contextual information, such as preceding comments or posts,Markov and Daelemans (2022) advocate for a more nuanced approach. They argue that merely adding previous comments or the content of a post does not necessarily enhance the performance of hate speech detection models. Instead, the key lies in identifying and integrating the relevant context. Their experiments demonstrate that when models focus on this relevant context, their performance in detecting the target of online hate speech improves significantly.

Markov and Daelemans (2022) emphasize that the effectiveness of hate speech detection models can be significantly enhanced by identifying and integrating the relevant context. In their experiments, using only the comment as a baseline, the model achieved metrics of precision at 0.65, recall at 0.66, and an F1-score of 0.63. When the content of the post was combined with the comment, there was a slight uptick in performance, registering a precision of 0.66, recall of 0.67, and an F1-score of 0.65. The addition of the preceding comment yielded metrics of precision at 0.64, recall at 0.65, and an F1-score of 0.63. And when both the preceding comment and the post content were integrated, the results mirrored the previous metrics. Most notably, when the model was tailored to focus on the relevant context, it showcased its highest performance, achieving a precision of 0.69, recall of 0.71, and an F1-score of 0.69.

The findings of Markov and Daelemans (2022) have profound implications for the development of hate speech detection models. The fact that these models can significantly improve their performance by focusing on the relevant context underscores the need for more research in this domain. The study concludes by emphasizing the potential of encoding contextual information and its significant effect on detecting the target of hate speech, suggesting promising directions for future research in this area.

2.7 Conclusion

The tension between the findings of Karan and Šnajder (2019) and Markov and Daelemans (2022) highlights a crucial gap in the research of toxic language detection. While the former suggests that the inclusion of thread-level information did not significantly boost the performance of context-sensitive models over context-agnostic ones, the latter argues that focusing on the relevant context can improve the performance of hate speech detection models in a significant way. This discrepancy indicates that there is more to uncover regarding the role of context in online toxic language detection. The optimal way of integrating context - be it the entire conversation thread, preceding comments, or other nuanced approaches - remains an area of interest for research.

Chapter 3

Dataset Characteristics

This chapter offers a detailed overview of the CAD by Vidgen et al. (2021). It investigates the approach that Vidgen et al. (2021) used to collect and transform Reddit posts into annotated data. Reddit is a platform consisting of an enormous array of communities called subreddits. Each subreddit typically deals, for example, with a particular topic or hobby in order for users to generate messages and response threads there. The discussion thread of each of these posts, consisting of one initial post followed by any number of comments below it and responses to these comments, is the primary source for CAD.

3.1 Data Collection

In their research, Vidgen et al. (2021) developed the CAD from 16 subreddits known for a wide range of abusive content. They initially screened 117 subreddits identified for toxic or offensive content. The selection criteria were: no political bias, no targeting specific groups, and recent activity. This process narrowed the field to 16 active communities such as r/Drama and r/conspiracy. Between February and July 2019, they collected 187,806 threads via the PushShift API. To make the data manageable, they narrowed this down to 1,394 posts and 23,762 comments, totalling 25,156 pieces.

This method addresses the scarcity of online abuse, which is only 0.001% of social media content (Vidgen et al., 2019). Traditional datasets often rely on targeted sampling to find abuse, which can bias the results and could undermine the performance of detection systems. Vidgen et al. (2021) chose a different route. Instead of relying on keywords or isolated comments, they used complete threads from selected subreddits. This approach ensures that both abusive and non-abusive entries have similar contexts with the aim to improve the accuracy and reliability of detection tools.

3.2 Annotation Process

The CAD annotation team was made up of 12 annotators, including both native English speakers and those fluent in English as a second language, who individually labeled the data according to the dataset's taxonomy (Vidgen et al., 2021).

3.2.1 Disagreements and Resolution

Despite the extensive training, the subjective nature of online abuse sometimes led to disagreements among annotators. These disputes often involved interpreting implicit abuse or understanding context-dependent posts. For instance, distinguishing sarcasm from genuine abuse or recognizing subtle language provided its own challenges. The team used a consensus-based approach to settle their differences.

3.2.2 Inter-Annotator Agreement (IAA)

To assess the consistency among annotators, Vidgen et al. used Fleiss' Kappa, a statistical measure of agreement among multiple raters. Fleiss' Kappa scores range from 0 (no agreement beyond chance) to 1 (perfect agreement). Scores above 0.6 indicate substantial agreement, while scores between 0.4 and 0.6 suggest moderate agreement.

Category	**Fleiss Kappa Score**
Non-hateful slurs	0.754
Neutral	0.579
Person-directed abuse	0.513
Affiliation-directed abuse	0.453
Identity-directed abuse	0.419
Counter Speech	0.267

Table 3.1: Fleiss Kappa Scores for Different Categories

According to table 3.1, the agreement scores between the annotators differed for each category, showing the varying levels of difficulty in labelling them. Non-hateful slurs saw the highest agreement, suggesting they were clearer to identify, whereas Counter Speech had the lowest score, making it the most difficult category to consistently label, probably due to its rare occurrence and the complexity involved in its context. While the CAD's moderate IAA scores are not ideal, they align with the challenges faced by other datasets in this domain. The CAD's overall Fleiss' Kappa score of 0.583 is comparable to other datasets in the field.Wulczyn et al. (2017) reported a score of 0.45 for their Wikipedia comments dataset, while Fortuna and Nunes (2018) reported a score of 0.58 for their Twitter dataset. This suggests that achieving high interannotator agreement is a common challenge due to the inherent subjectivity of online abuse.

3.3 Annotation Overview

In this section, I will provides an overview of the annotation guidelines and the taxonomy that are used in collecting the CAD dataset.

3.3.1 Annotation Guidelines

The annotation guidelines were developed to ensure a reliable approach to labelling the dataset. Annotators underwent a four-week training period. They received detailed instructions on how to apply the taxonomy to real-world examples from Reddit conversations. Training included group discussions and individual feedback sessions to ensure that the annotators understood the categories thoroughly and could apply them consistently. Annotators were told to consider the entire conversation thread when interpreting posts. The guidelines stressed the importance of distinguishing between overt and covert forms of abuse. This includes identifying subtle cues and implicit language that could indicate abusive content. When faced with challenging or ambiguous cases, annotators consulted with the team, and regular group discussions were held to review difficult examples. The guidelines were refined based on the annotators' experiences.

3.3.2 Taxonomy and Definitions

CAD's taxonomy consists of six primary categories:

Neutral: Content that does not contain abuse or is unrelated to the topic of abuse.

Example: "I've had a right bloody day of it"

Identity-directed Abuse: Negative statements targeting an individual's or group's inherent characteristics, such as race, gender, sexuality, or disability. This category captures content that explicitly or implicitly demeans, insults, or attacks individuals or groups based on these inherent traits.

Example:

Parent text: "Someone who doesn't think you need dysphoria to be transgender"

Reply: "Oof. That's dumb. Makes actaul people with dysphoria look dumb..."

Affiliation-directed Abuse: Abuse targeting someone's political, ideological, or other chosen affiliations, such as their membership in an organisation or their beliefs and opinions. This category includes attacks or insults based on these affiliations.

Example:

Parent text: "Rappers usually get a pass."

Text: "Only when it comes to selling drugs and murder, the left will devour them if they attack one of their precious minorities."

Person-directed Abuse: Direct attacks or insults towards specific individuals, distinguished from identity-directed abuse by its personal nature. This category includes targeted harassment, name calling, and other forms of personal attacks that are not necessarily tied to the target's identity or affiliations.

Example: "What a fucking idiot. Honestly ... when you have to reach that hard there's got to be something broken in your head."

Counter-Speech: Responses that challenge or call out abusive language or behavior. The goal of these posts is to promote a more respectful and inclusive discourse. This category includes comments that dispute or criticise abusive remarks, educate others about the harmful nature of a certain language, or support targeted individuals or groups.

Example:

Parent text 1: "Slut has always been an insult for women, not men, and virgin has always been an insult for men, not women. This whole dichotomy exists because of the different ways in which men and women pursue sex. It is not difficult to attract sexual partners as a woman. The same is not true for men. Attracting partners is actually challenging as a man, so the ability to do so is seen as that - an ability, a talent, a skill. Now, a woman who has high standards will find difficulty attracting a man who fits them, but it is fundamentally not hard for a woman to find sex if that's all she wants, so choosing whether to be a slut or not is always a conscious choice. If she wants sex, she can get it. If she doesn't, she doesn't. Right-wing society says the choice to be a virgin is the superior one, left-wing society disagrees, but it's still a choice. For men, attracting partners is not a choice, it is a challenge. Thus, calling a man a virgin is not at all like calling a woman a slut.”

Parent text 2: “well said.”

Reply: “that’s not well said. it’s just trying to justify that it’s okay to attack people using sexuality as long as you have the right politics.”

Non-hateful Slurs: The use of derogatory language that has been reclaimed by the targeted group or does not carry hateful intent in the given context. This category includes instances where slurs or offensive terms are used in a non-hateful way, such as in-group reclamation or when the context clearly shows no harmful intent.

Example:

Parent text: “I was gonna say ”“it was probably just a typo”“ but nope, they literally replaced every ’n’ with an ’m’”

Reply: “probably they don’t know what the n-word is but they don’t want to say it. so to avoid saying the n-word they just remove all Ns from words”

3.3.3 The Role of Context

During the annotation process of CAD, annotators were instructed to consider the context provided by previous posts when labeling comments. This contextual information was crucial in accurately identifying instances of toxic language, as the meaning and intent of a comment can often be influenced by the surrounding conversation.

It’s important to note that context was used sparingly in the annotation process, with only 6.38% of comments requiring contextual information for proper interpretation. This indicates that the majority of comments could be accurately labeled based on their content alone. Additionally, in some cases where context was utilized, it was primarily for coreference resolution rather than directly determining whether the comment was toxic or not.

In the CAD, the “annotation_Context” column indicates whether the annotation of a comment relied on context from previous posts. If the value is “PreviousContent,” it signifies that the annotators used contextual information to make their decision, whereas “CurrentContent” suggests that the annotation was based solely on the content of the current comment.

By using Annotation Context and reconstructing the chain of previous comments, it can be seen how contextual information can significantly influence the interpretation of a comment and reveal nuances that might be missed when examining the text in isolation. However, this contextual analysis was only necessary for a small portion of the dataset.

Consider the following example from CAD (cad_3462): “Name a time where a woman deceived someone. I’ll wait.”

Label: IdentityDirectedAbuse

In isolation, the text might not be considered as abuse that is directed towards woman. However, when read in the context of its parent comments, it becomes clear that this is not a genuine question but rather a sarcastic response to the previous comments discussing the credibility of women’s accusations. The following is the full context:

Parent Text 1: “Really, ”“Believe All Victims”“ needs to come with an asterisk after ”“Believe”“. I agree that you shouldn’t respond to accusations with a knee-jerk ”“you’re a liar”“, you should offer sympathy and should lean towards believing them, since the vast majority of accusations (98% of all crimes and 91% of sexual crimes, according to the FBI IIRC) are true. What you should also do is weigh all the evidence as it is made known and if the evidence suggests that the crime was faked or the accusation is false to call that out as soon as is reasonable. Trust, but verify."

Parent Text 2: "Here's where I think the distinction lies - I may be wrong, and call me out if you want, but here goes: So, believe all victims is a good and noble idea at a basic level. When a victim of a crime comes forward, the people that they trust should absolutely support them, and help them get help and try to make it as easy as possible to get the police involved - not necessarily influence them to go to the authorities, but that's where crimes should go, for a number of reasons. This becomes a less tenable position when a) the justice system gets involved and b) when there are real world consequences for leveling an accusation. With the Jussie Smollett case, we see both of those in play. When the police investigation unfolded, they were able to put forth compelling evidence against his claim. The second set is more nuanced, but this is how I see it: His claims had a major effect to divide the country. Seeing these claims would have such an effect, it gives us good reason to scrutinize his story. When we scrutinize the story, we find some major holes (two maga people out after 2 in intensely cold weather, in Chicago, apparently knowing where Smollett was, and being prepared for that confrontation, his refusal to hand over his phone, he walked home and waited 40 minutes or so before calling the police) and we have a good reason to step back a bit from the narrative and wait for an investigation to conclude before deciding. In short, believe all victims is a good idea, but it becomes more nuanced when there are real consequences that may hurt or jail others, at which time we should scrutinize the story or let the professionals do their job."

Parent Text 3: "#BelieveEveryBullshitStory"

Parent Text 4: "and here i thought you were more than the weakest of trolls..."

Parent Text 5: "Women can't deceive people though."

Parent Text 6: "oh boy, treating women like they are capable of deceit sure is hating on them. Wouldn't want to think they could do everything a man can do but better now would we?"

Text: "Name a time where a woman deceived someone. I'll wait."

Label: IdentityDirectedAbuse

Only by situating this comment in the full thread, it becomes clear that this comment is part of the pattern of misogynistic views and dismissal of women's experiences becomes evident. Another example which can reveal the importance of context is (cad_3816). Text: "Despite making up only 13% of the population."

Label: IdentityDirectedAbuse

Again, the text in isolation is not indicative of a form of abuse, however, within context, the reasoning for this categorization becomes clear. The following is the full context:

Parent Text 1: "Martin Luther King was a Freedom loving Christian!"

Parent Text 2: "Well it's true. Whites and blacks are the same. Men and women aren't. Only differences between white and black are cosmetic

Edit: oof I guess vaccines also cause autism"

Text: "Despite making up only 13% of the population."

Label: IdentityDirectedAbuse

When viewed in isolation, the text "Despite making up only 13% of the population" might seem incomplete or ambiguous. However, when placed in the context of the parent texts discussing racial differences and making controversial claims, it becomes evident that the comment is likely referencing a racist trope about black people and crime statistics. Within the full thread, the conversation reveals a pattern of racist views that is being perpetuated to reinforce harmful stereotypes about a minority group in the United States.

3.4 Dataset Composition and Distribution

The CAD is a collection of 27,494 entries with annotations that includes both abusive and non-abusive content (Vidgen et al., 2021). After processing, the dataset is divided into training (58%), development (19.3%), and test (22.7%) sets (Vidgen et al., 2021), with a total of 23,417 eligible entries. The distribution of the primary annotations is presented in tables 3.2 and 3.3.

Category	**Percentage**
Neutral	79.78%
Identity Abuse	9.93%
Affiliation Abuse	4.91%
Person Abuse	4.04%
Counter Speech	0.80%
Slur	0.54%

Table 3.2: annotation Statistics

The dataset exhibits a significant class imbalance, with most entries labelled neutral (79.78%). This imbalance aligns with the distribution of abusive content in real-world social media conversations, where most interactions are not abusive (Fortuna and Nunes, 2018; Zampieri et al., 2019), Among the categories of abusive content, identity abuse is the most prevalent with 9.93%, followed by person abuse (4.91%) and affiliation abuse (4.04%). Counter Speech and Slurs are the least represented categories, comprising only 0.80% and 0.54% of the dataset, respectively.

Category	**Dev**	**Test**	**Train**
Neutral	78.22%	78.71%	78.01%
Identity Abuse	10.10%	10.74%	10.80%
Affiliation Abuse	5.79%	4.70%	5.44%
Person Abuse	4.40%	4.31%	4.34%
Counter Speech	0.81%	1.20%	0.80%
Slur	0.68%	0.35%	0.60%
Total Comments	4,684	5,495	14,113
Percentage	17.04%	19.99%	51.33%

Table 3.3: Distribution of primary annotations and comments across datasets

3.5 Thread Dynamics Analysis

In this section, I will draw from my analysis to examine the average thread lengths and word counts for each entry. This analysis is important because BERT has a max length limit of 512 tokens. Moreover, the longer the thread takes, the previous comments might lose their relevance to the comment being analyzed and could potentially introduce noise. My analysis shows that the average number of preceding comments for a given comment in the CAD is 3.92. Furthermore, examining the distribution of the number of preceding comments per thread reveals that, for 90% of the comments, there are 4 or fewer preceding comments. The word count analysis also shows the thread length, measured by the total word count of the initial post and all preceding comments (Figure 3.1), and the distribution of thread lengths in the CAD. This analysis indicates that most threads are relatively short:

- 80th percentile: 118 words
- 90th percentile: 185 words
- 95th percentile: 260 words

This means that most of the threads in the dataset contain a manageable amount of contextual information. However, there are some longer threads, with a few exceeding 1,000 words in total length. The short thread lengths and low number of preceding comments for most of the data suggest that the majority of comments have a manageable amount of context. But the existence of some longer threads means there needs to strategies to handle input constraints when applying certain models to the dataset such as truncating the input which will be discussed in the next chapter.

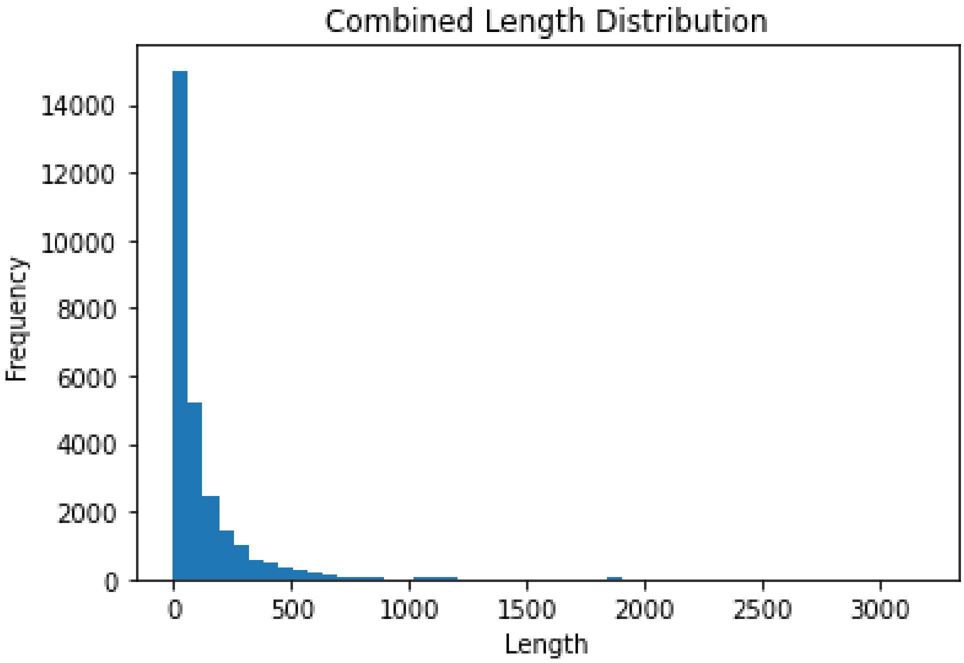

Figure 3.1: Combined length thread distribution

3.6 Content Analysis

After preprocessing the text data including tokenization, stop word removal, and non-content phrase filtering, analysis of the most frequent terms provides some linguistic insights into how hate speech and non-hate speech conversations differ stylistically. It is important to note that a different preprocessing approach was used for the classification tasks discussed in the next chapter parts of this thesis. The most common words in hate speech threads highlight tendencies towards increased profanity and references to identity. Frequent terms include “fucking” (633 occurrences), “shit” (463), “fuck” (445), “white” (376), and “women” (350). Additionally, political terms like “left” (283), gendered words such as “men” (271), and emotional words including “hate” (229) appear often.

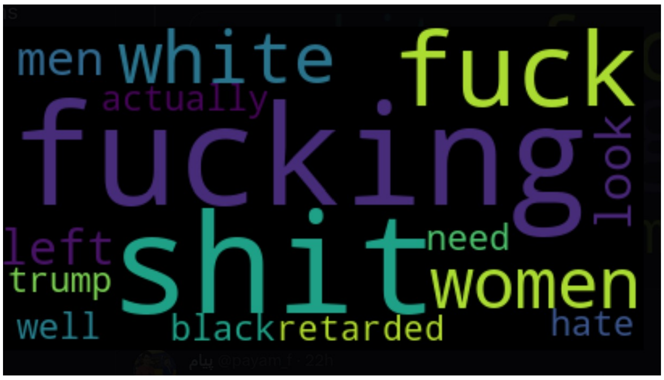

Figure 3.2: word cloud for toxic category

In contrast, top words in non-hate speech threads tend to be more neutral and polite. Common terms include general filler words like "well" (675) and "actually" (632), polite phrases such as "please" (578), and conversational words like "said" (551). While there is some overlap in frequently used words across both categories, such as "shit" (696 in non-hate speech, 463 in hate speech), "well" (675 in non-hate speech, 247 in hate speech), and "fucking" (457 in non-hate speech, 633 in hate speech), the non-hate speech conversations generally lack the strong profanity and identity focus observed in hate speech threads.

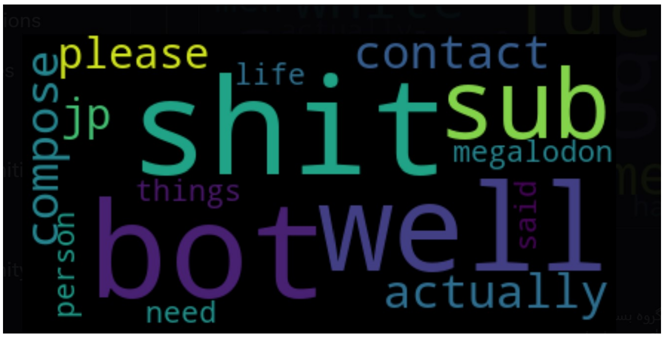

Figure 3.3: word cloud for non-toxic category

This comparative frequency analysis reveals stylistic differences between these categories of threads. However, it does not capture nuanced contextual usage of terms. Further discourse analysis may provide additional insights into the linguistic patterns differentiating hate and non-hate speech.

3.7 User Analysis

I analyzed the dataset in terms of user activity and the prevalence of toxic language to gain an understanding of the characteristics that might potentially give us a useful addition for automatic classification. In the context of this analysis, users are defined as the unique authors of the entries in the dataset, which are identified by their Reddit usernames in CAD.

To examine the relationship between user activity and the likelihood of posting toxic language, I calculated the correlation between the number of posts per user and the share of their posts that were marked as containing (Identity-directed, affiliation-directed, or Person-directed). 11,123 distinct users can be found in the dataset. The resulting correlation coefficient is 0.0103, indicating a very weak positive relationship. This suggests that more active users have a slightly higher tendency to post toxic language, but the effect is minimal.

It is important to note that this distribution may not necessarily reflect the natural prevalence of toxic language among Reddit users, as the dataset was purposefully sampled from subreddits known to contain higher levels of abusive content (Vidgen et al., 2021). As such, the weak correlation between activity and toxic language could be a result of the dataset's selection process.

Examining the distribution of mean toxic language rates per user (Figure 3.4) reveals a largely left-skewed distribution, with the majority of users having no toxic posts. However, there is also a notable spike on the far right, representing a small subset of users with 100% toxic language rates. This suggests that a few users are responsible for a disproportionate amount of toxic content in the dataset.

Furthermore, I established a cutoff point: users whose posts contain 50% or more toxic language are considered "high-toxicity" users. Applying this standard, 1,475 users, or 13% of the total examined, were categorised as high toxicity. This observation suggests that a relatively small group of users in this dataset is responsible for a significant portion of the toxic language.

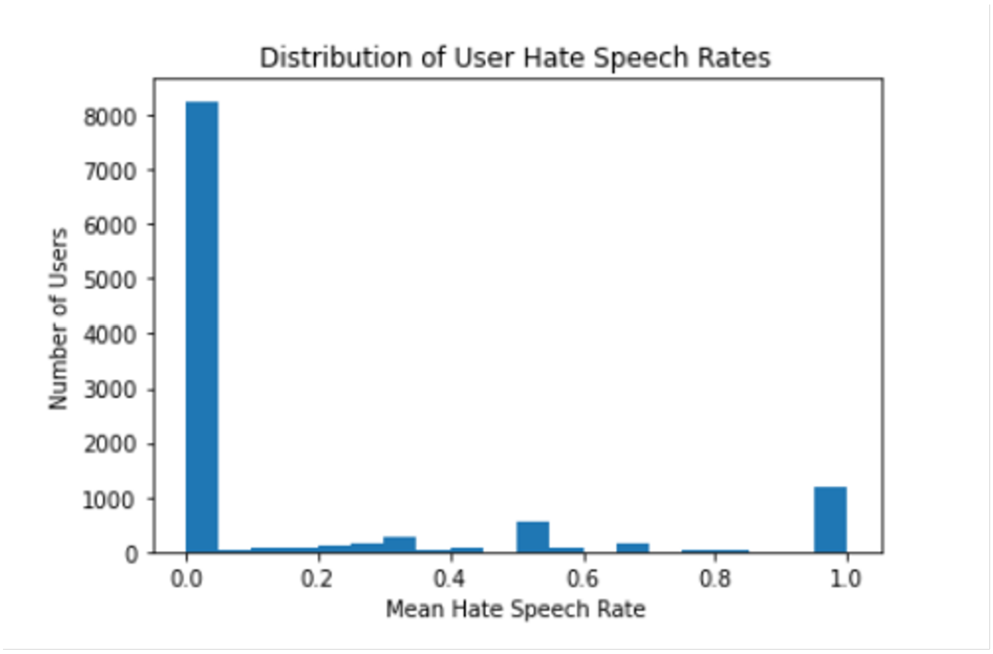

Figure 3.4: Mean Hate Speech Rate

3.8 Temporal Analysis of Hate Speech

In my analysis of hate speech prevalence over time in the CAD, I observed that the rates fluctuated from week to week (Figure 3.3). The lowest proportion of hate speech was in week 2019-04 (16.52%), which corresponds to April 2019, while the highest was in week 2019-06 (24.08%) in February 2019.

Taking a closer look (figure 3.5), I found that April (weeks 2019-14 to 2019-17) and May (weeks 2019-18 to 2019-22) had the lowest proportion of hate speech at only 16-17% of posts. On the other hand, June (weeks 2019-23 to 2019-26), July (weeks 2019-27 to 2019-30), September (weeks 2019-36 to 2019-39), and December (weeks 2019-49 to 2019-52) had higher proportions, ranging from 22-24%. While these observations hint at a potential seasonal pattern, with higher rates in early summer and late fall/early winter, it's important to keep in mind that the dataset's selection process could have influenced these distributions.

I also analyzed the distribution of hate speech posts by hour (Figure 3.6) and found that it varies throughout a typical day, from a low of 10.5% to a peak of 22.3%. There are notable spikes around 11am (22.3%), as well as in the early morning and late afternoon around 6-7am and 4-5pm (both 20.9%). The early evening around 7-8pm also has an elevated percentage of 21.8%. In contrast, the percentage drops to its lowest point of 10.5% during the late night around midnight and reaches another low of 18.5% at 4am.

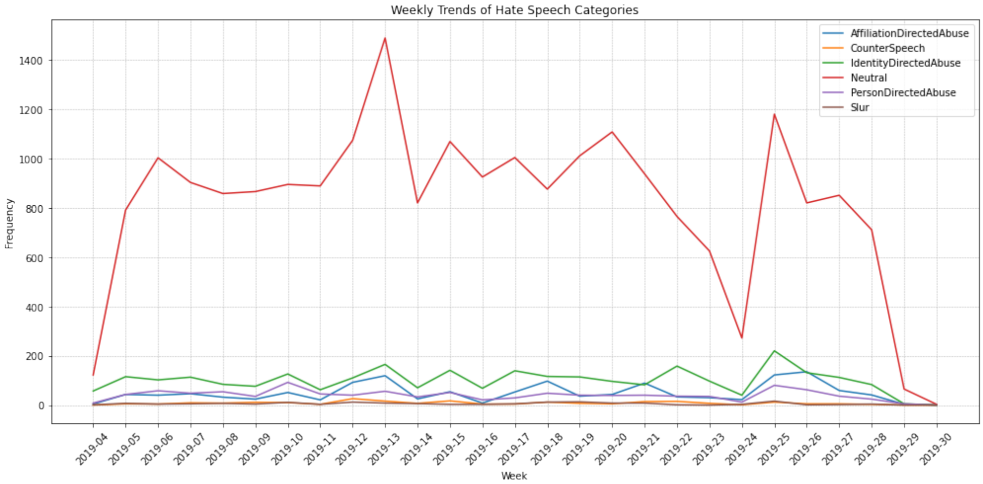

Figure 3.5: Weekly Trends of User Activity

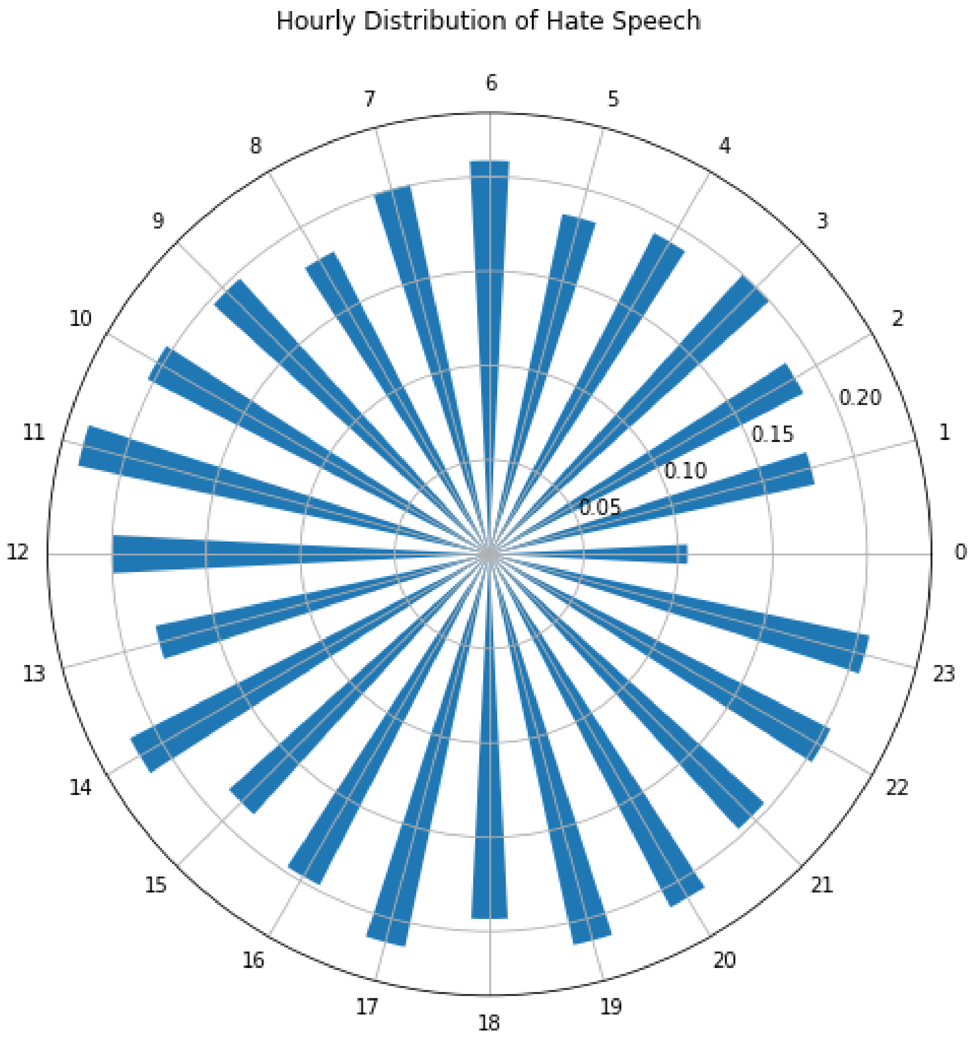

Figure 3.6: Hourly Distribution of Hate speech

3.9 Conclusion

This chapter provided an overview of CAD characteristics through conducting various analyses. One of the main features of CAD discussed in this chapter was the imbalance of the classes in the dataset, with 79.78% of entries labeled as Neutral and 20.22% constituting various abusive categories.

Furthermore, the analysis shows that the annotators only used the context from previous comments 6% of the time and in the remaining 94%, the annotators were able to base their labels on the comment that was being classified alone. This suggests that most comments in CAD can be understood on their own without a need to rely on the context.

In terms of thread lengths, most conversations in CAD are relatively short, with 90% having 4 or fewer preceding comments. Additionally, 80% of threads total 118 words or less when combining all comments. However, some longer outlier threads do exist that exceed typical model input limits, highlighting the need for strategies like truncating to handle these cases. The thread lengths and The above-mentioned qualities are the most important characteristics of this dataset with regard to developing a toxic language detection model.

Chapter 4

Methodology

4.1 Overview of BERT

BERT (Bidirectional Encoder Representations from Transformers) emerged as a state-of-the-art natural language processing model in 2019 (Devlin et al., 2018). Unlike unidirectional models before it, BERT leverages a bidirectional Transformer encoder architecture to learn contextual relationships between words, providing an advantage in understanding conversational language. This bidirectional conditioning is enabled through a tailored, two-phase training procedure; an initial pretraining stage followed by task-specific fine-tuning.

In pretraining, BERT learns bidirectional representations from scratch on unlabeled data through two unsupervised prediction tasks. First, in Masked LM random tokens are masked and predicted based on context. Second, Next Sentence Prediction predicts order relationships between sentences. After pretraining, BERT is fine-tuned on labeled data by adding a classification layer and training the pretrained weights on this downstream task data (Devlin et al., 2018). All parameters are tuned end-to-end using backpropagation to maximize performance on the specific task.This two-step approach allows BERT to first build extensive linguistic knowledge through pretraining, then specialize for a particular task through fine-tuning (Merchant et al., 2020).

Although BERT's architectural innovations facilitate the modeling of linguistic context, the effective processing of raw text input remains a critical aspect of its implementation. This is accomplished through a tokenization process, specifically WordPiece tokenization, which enables BERT to mitigate out-of-vocabulary issues and process any given raw text input. BERT employs WordPiece tokenization where rare or unknown words are split into meaningful subword units (Devlin et al., 2018). Additionally, special classification ([CLS]) and separation ([SEP]) tokens are inserted. The [CLS] token is added to the start of every sequence to represent the entire sequence. Meanwhile, [SEP] tokens separate different sentences or sentence pairs within a sequence.

BERT relies on a multi-layer Transformer encoder architecture for modeling linguistic context. The Transformer blocks provide BERT the capacity to jointly incorporate contextual information from both directions. At its core, BERT utilizes stacked Transformer encoder blocks rather than recurrent networks like LSTMs (Devlin et al., 2018). Each Transformer block contains two main components - a multi-head self-attention layer and a position-wise feedforward layer .

A distinguishing component of the Transformer architecture is the multi-head self-attention mechanism, which allows BERT to attend to different positions in the input sequence concurrently. This enables modeling of linguistic context by determining the relevance of each token to the rest of the sequence. Technically, self-attention consists of parallel attention layers focused on distinct aspects of the input. Through attending to particular positional signals, the individual attention heads enable BERT to simultaneously interpret multiple contextual relationships. The attention weights indicate the relevance of each input sequence for generating the contextual output representation. Essentially, multi-headed attention provides a key capability for BERT to learn which parts of the input are most relevant for encoding contextual relationships. The concatenated outputs from the multiple heads provide fine-grained attention across positions. The concurrent attention to multiple positions enables the Transformer architecture of BERT to achieve substantial performance improvements on language tasks compared to recurrent networks (Vaswani et al., 2017).

4.2 Bert for Hate Speech detection

As discussed, BERT's architecture relies on a bidirectional approach to process text. This indicates that the interpretation of each word is influenced by both the preceding and following words in a sentence, unlike older models that only considered one direction of context. Such contextual understanding is critical in hate speech detection, where the meaning and intent behind words can vary significantly based on their linguistic environment (Gao and Huang, 2017). For example, certain phrases or words might be innocuous in one context but could convey hate speech in another. BERT's ability to detect these subtleties makes it well-suited for this task.

Moreover, BERT's training process provides it with a nuanced understanding of language (Kishimoto et al., 2020). This is especially beneficial in identifying hate speech. The model, through its extensive pretraining, develops a foundational understanding of language, which is then refined to the specifics of hate speech detection during the fine-tuning phase. This process supposes that BERT is not just recognizing blatant instances of hate speech but is also sensitive to more subtle forms.

Furthermore, BERT's employment of the Transformer architecture allows it to weigh and integrate different parts of a sentence or text snippet to form a more comprehensive understanding. This is an important aspect in analyzing conversations in the CAD, where in some comments the context and sequence of messages play a significant role in determining whether a statement is harmful or not (Vidgen et al., 2021).

The Transformer architecture employed by BERT is significant in its ability to comprehend and analyze hate speech within the CAD. This architecture allows BERT to effectively weigh and integrate different components of a sentence or text snippet, forming a more comprehensive understanding of the content. In the context of the CAD, where the sequence and context of messages could play a significant role in determining the harmful nature of a statement, this capability is particularly valuable. By considering the contextual relationships between words and sentences, BERT can better identify and interpret the subtle linguistic cues that are characteristic of toxic language.

4.3 Research Approach and Experimental Design

This research uses a multifaceted experimental design to test how well BERT can detect toxic language in the CAD. The experimental approach is twofold, focusing on different context levels and comparative analyses across various categorizations of the dataset.

4.3.1 Context levels

Three distinct BERT-based models are developed, each differing in the scope of textual context considered:

1. **Comment-Only Model**: This model focuses solely on the individual comments that are labeled, disregarding any surrounding conversational context. The aim is to ascertain the effectiveness of BERT in identifying hate speech based purely on the content of a single comment.

2. **Comment with Parent Model**: The second model extends the analysis to include both the labeled comment and its immediate parent or preceding comment. This design is intended to explore the impact of immediate conversational context on the detection accuracy.

3. **Full Thread Model**: The third model incorporates the entire conversation thread leading up to the labeled comment. This approach aims to evaluate how broader context spanning multiple conversational exchanges influences the model's ability to detect hate speech.

4.3.2 Comparative Analyses

This research uses a comparative methodology, which contrasts different model configurations and categorization approaches. By comparing the outcomes across these varied setups, the study aims to gain insights into the strengths, limitations, and optimal applications of BERT in the context of hate speech detection. Two sets of comparisons were conducted to understand the nuances of hate speech detection:

1. **multiple class Analysis**: Based on the original design of the dataset, there were six categories which were used to classify comments: Neutral, Identity Directed Abuse, Affiliation Directed Abuse, Person Directed Abuse, Non-hateful slurs, and Counter Speech (as defined in section 3.3.2). Following the original study, Non-hateful slurs and also Counter Speech were categorized as Neutral due to their low occurrence in the dataset.

2. **Binary Classification Analysis**: To simplify the classification, a binary scheme was adopted, collapsing all categories into two: "Neutral" and "Abusive." Given the disparity in the distribution of these two classes, undersampling was employed to balance the dataset, reducing the "Neutral" category from 14,825 instances to match the 3,285 "Abusive" instances.

4.3.3 Experimental Design

The experimental design aims to test the capabilities and limitations of BERT across various contextual settings for toxic language detection. Six distinct experiments are conducted, each exploring different contextual settings and classification approaches.

Experiment 1: Comment-Only Model (Binary Classification)

This experiment focuses solely on individual comments, disregarding any surrounding conversational context. Its objective is to evaluate BERT's effectiveness in identifying toxic language in a binary classification setup, categorizing comments as either "Neutral" or "Abusive."

Experiment 2: Comment-Only Model (Multiclass Classification)

Similar to Experiment 1, this variant classifies comments into multiple categories based on the dataset's taxonomy, which includes "Neutral," "Identity Directed Abuse," "Affiliation Directed Abuse," and "Person Directed Abuse."

Experiment 3: Comment with Parent Model (Binary Classification)

Extending the analysis, this experiment includes both the labeled comment and its immediate parent or preceding comment. Its goal is to assess the impact of immediate conversational context on toxicity detection in a binary classification setup, categorizing comments as either "Neutral" or "Abusive."

Experiment 4: Comment with Parent Model (Multiclass Classification)

Similar to Experiment 3, this variant utilizes multiclass classification, categorizing comments into "Neutral," "Identity Directed Abuse," "Affiliation Directed Abuse," and "Person Directed Abuse."

Experiment 5: Full Thread Model (Binary Classification)

Incorporating the entire conversation thread leading up to the labeled comment, this experiment evaluates how broader context influences toxicity detection in a binary classification setup, categorizing comments as either "Neutral" or "Abusive."

Experiment 6: Full Thread Model (Multiclass Classification)

Similar to Experiment 5, this variant employs multiclass classification, categorizing comments into "Neutral," "Identity Directed Abuse," "Affiliation Directed Abuse," and "Person Directed Abuse."

By conducting these six experiments, the study aims to test BERT's performance in detecting toxic language across different contextual settings (comment-only, comment with parent, and full thread) and classification approaches (binary and multiclass).

4.4 Preprocessing CAD for BERT

Several key adjustments were made to CAD before processing with training. The primary goal was to align the dataset with the input requirements of BERT, focusing on tokenization and handling of special tokens. The Python script utilized for this purpose encompassed a range of preprocessing tasks, ensuring the dataset was optimally formatted for effective analysis by BERT.

4.4.1 Creating Hierarchical Datasets

One of the significant contributions of this research is the transformation of CAD into a hierarchical structure that includes parent comments. I transformed this dataset by developing a custom code specifically for this purpose which is available at [the code repository: https://github.com/pafa3/context-aware-CAD]. The goal of this modification was to investigate the impact of conversational context on hate speech detection by incorporating the entire conversational hierarchy into the dataset.

The original CAD comprised Reddit posts and comments as separate entries. To enhance the dataset, each comment was associated with its respective parent comments, extending up to 15 levels, corresponding to the longest thread in the dataset. The creation of the new dataset involved several key steps:

First, Identifying the parent comment for each entry by referencing the 'info_id.parent' field that was originally in the dataset.

Second, recursively extracting parent information by tracing back through the comment thread for each comment, gathering both text and author details from each parent comment level.

Third, modifying the dataset to include new columns that accommodate the hierarchical structure.

Table 4.1 presents an example that illustrates the transformed dataset. For a given comment, the 'meta_author' and 'meta_text' columns contain the author and text of the comment itself. The 'parent_text_level_n' and 'parent_user_level_n' columns represent the text and author of the parent comment at each level, respectively. This hierarchical structure enables a comprehensive analysis of each comment within its broader conversational context.

meta_author	Awayfone
meta_text	Dont be jelous
parent_text_level_0	Don't forget the BBC she enjoys taking on the reg
parent_user_level_0	egadsby
parent_text_level_1	The only black Demi likes is the black tar heroin giving her that sweet sweet high
parent_user_level_1	TheTrueNobody
parent_text_level_2	Demi Lovato tweets out the most innocuous meme imaginable prompting the "that's problematic" lynch mob to bully her into deleting her account
parent_user_level_2	Strictlybutters

Table 4.1: Example of the hierarchical structure in the transformed dataset (transposed)

By incorporating parent comments, the dataset now provides a more comprehensive view of the conversational dynamics that shape the interpretation and impact of each comment. Moreover, the inclusion of author information at each level of the hierarchy allows for the examination of user-level patterns and behaviors, which could provide insights into the propagation of hate speech within online communities.

4.5 Data Preparation for Hierarchical Context Analysis With BERT

After restructuring CAD to include hierarchical conversational context, further data preparation was necessary to ensure compatibility with BERT's framework. This preparation involved several key steps to refine the dataset for nuanced hate speech detection.

4.5.1 Data Consolidation and Level Extraction

The dataset entries underwent a consolidation process, preparing them for analysis with BERT. This preparation varied depending on the desired level of contextual depth, ranging from level 1 to level 3.

- **Level 1**: At this level, the focus was solely on the comment text itself, without including any parent comment context.
- **Level 2**: This level expanded the analysis to include the immediate parent or preceding comment, providing direct conversational context.
- **Level 3**: The most comprehensive level involved concatenating text from up to 15 preceding comments. Individual comments were separated by '[SEP]' which serves as a signal to BERT to identify different segments of the thread.

An illustrative example from the training dataset (entry 297) demonstrates this structure:

Level 1:
Text: Dont be jelous
Labels Info: 1 (IdentityDirectedAbuse)
Parent Text:
Level 2:
Text: Dont be jelous [SEP]
Parent Text: Don't forget the BBC she enjoys taking on the reg [SEP]
Level 3:
Text: Dont be jelous [SEP]
Parent Text: Don't forget the BBC she enjoys taking on the reg [SEP]

The only black Demi likes is the black tar heroin giving her that sweet sweet high [SEP]

Demi Lovato tweets out the most innocuous meme imaginable prompting the "that's problematic" lynch mob to bully her into deleting her account

Labels Info: 1 (IdentityDirectedAbuse)

It is important to note that while the labels in the dataset are associated with the comment text stored in the 'text' field, the BERT model actually learns to make predictions based on the combined input of both the 'text' field (the comment itself) and the 'parent_text' field (the preceding comments in the thread) during the training and inference process In the data preprocessing stage, the getitem method concatenates the 'text' and 'parent_text' fields and encodes them together using the tokenizer. Therefore, when the BERT model processes this sample, it takes into account the entire input sequence, including both the 'text' and 'parent_text'. The model learns to associate the label with the combined input, not just the 'text' field alone. Therefore, it is crucial to point out that although the labels are originally assigned to the 'text' field in the dataset, the BERT model's predictions are based on the combined input of 'text' and 'parent_text'.

4.5.2 Text and URL Substitution

To focus the analysis on the linguistic content of each comment, specific textual elements in the dataset were replaced with generic placeholders. Subreddit references and usernames were substituted with '[subreddit]' and '[user]' tokens, respectively. Similarly, URLs were replaced with a '[LINK]' token. These substitutions were performed to prevent the model's focus on potentially irrelevant details, ensuring a concentration on the textual content.

Furthermore, entries that were irrelevant or lacked substantive content, such as those marked as "[removed]" or "[deleted]," were excluded from the dataset. This step ensured the quality and relevance of the data being fed into the model

4.5.3 Label Encoding

The CAD's varied categories of speech, such as "Neutral," "IdentityDirectedAbuse," and others, were encoded into numerical labels. This encoding was essential for BERT's processing, as it requires numerical input for classification tasks. A mapping from textual labels to integers was established and applied across the dataset, facilitating the model's interpretation of the label data.

4.6 Fine-Tuning BERT

The fine-tuning of BERT for this research involved several steps designed to optimize the model's performance in detecting hate speech within the CAD. These steps ensured that BERT could effectively process and analyze the dataset, taking into account its unique characteristics and the specific requirements of the task.

4.6.1 Limiting Token Length

In the fine-tuning process of BERT for the CAD, a key consideration was determining the appropriate token length. To address computational limitations and facilitate multiple training iterations, the upper 95th percentile of word counts in the combined text (including parent text for levels 3) was set as the token length limit. This approach ensured that the majority of the conversational context was captured within the computational constraints, providing a practical yet effective basis for model training. This decision was crucial in optimizing the BERT model for efficiently processing the dataset while retaining essential contextual information.

4.6.2 Loss Function for Multi-class Classification

In addressing the multi-class classification task within the CAD, the Cross-Entropy Loss (CELoss) function was employed, which is a common choice for multi-class classification problems. The CELoss function combines a softmax activation with a negative log-likelihood loss, effectively handling the classification of samples into multiple mutually exclusive classes (Verbrugge, 2021).

Given the inherent class imbalance in the dataset, it was crucial to adapt the loss function to this problem. In standard settings, class weights are equal by default. However, to counter the imbalance observed in the CAD, weights were applied to the losses. This approach involved calculating the fractions for each class, resulting in a weight vector that accurately reflects the actual distribution of classes in the dataset. Such weighting ensures that the loss function accounts for the disproportionate representation of categories, thereby facilitating a more balanced training process.

4.6.3 Hyperparameter Optimization

In the fine-tuning phase of BERT, a critical aspect was optimizing the hyperparameters, particularly the learning rates, for each level of contextual analysis. The original train-test-validation split from the CAD dataset was maintained, and the validation set played a crucial role in this optimization process. The validation set, which consists of 19.3% of the total dataset as it was stated in chapter 3, was used to evaluate the model's performance during training and to guide the selection of the best hyperparameters.

In the fine-tuning phase of BERT, a critical aspect was optimizing the learning rates for each level of contextual analysis. The focus was not solely on performance metrics like accuracy or F1 score but primarily on the consistent decrease in validation loss. This approach was chosen as it provides a more reliable indicator of the model's ability to generalize, rather than just memorize the training data.

Chapter 5

Results and Findings

5.1 Introduction

This chapter will provide a comprehensive evaluation of the models used for detecting toxic language in CAD. It seeks to answer the main research question regarding the effect of progressively adding context—from no context to the entire conversation thread—to the BERT model's ability to detect and categorize abusive content. Focusing on three experimental conditions: analyzing only the comment, combining the comment with its preceding parent comment, and incorporating the entire conversation thread, this chapter will present and discuss the classification reports from each model.

5.2 Model Performance Overview

	Without Context			With One Parent			Full Thread		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Classification Report for multi-class classifiers									
Accuracy			0.81			0.81			0.80
Macro avg	0.49	0.52	0.49	0.49	0.49	0.49	0.48	0.51	0.49
Weigh. avg	0.82	0.81	0.81	0.81	0.81	0.81	0.81	0.80	0.81
Classification Report for binary-label classifiers									
Accuracy			0.78			0.74			0.74
Macro avg	0.66	0.73	0.68	0.65	0.73	0.66	0.65	0.73	0.65
Weigh. avg	0.83	0.78	0.80	0.83	0.74	0.77	0.83	0.74	0.76

Table 5.1: Classification Performance Metrics for Multi-class and Binary-label Classifiers

As Table 5.1 illustrates, in the multi-class model, the introduction of context—be it a single parent comment or the full thread—does not significantly improve the model's accuracy in identifying toxic language. The accuracy remains largely consistent, with an 81% accuracy rate when no context is provided and with one parent comment, only marginally decreasing to 80% with the full thread. However, in the experimental setup where we combined the labels for classifiers, the addition of context has introduced a

Support (Multi-class)			
Category	Without Context	With One Parent	Full Thread
Neutral	4407	4403	4410
AffiliationAbuse	205	236	241
IdentityAbuse	514	437	437
PersonAbuse	181	231	219

Support (Binary-label)			
Category	Without Context	With One Parent	Full Thread
Neutral	4410	4401	4401
Abusive Speech	897	906	906

Table 5.2: Support values for Multi-class and Binary-label Classifiers

level of complexity that affected the binary classifiers with context for the worse.

The findings indicate a need to explore the classifiers' performance in more detail to understand where the models excel and where they fall short. This will help us identify situations where context is most informative and develop strategies for effectively leveraging context to improve the detection of abusive language.

5.2.1 Multi-Class Classification Models

In examining the multi-class classification models' performance, the table 5.3 indicates that all models maintain high precision and recall in identifying 'Neutral' content. Notably, the precision marginally fluctuates between 90.20% and 90.94% across the models, suggesting that the presence or absence of context does not significantly impact the models' ability to recognize non-abusive content. However, the Model With One Parent Context shows a slightly better balance between precision and recall, achieving the highest F1-score at 90.24%, indicating a slight edge in overall performance for neutral classification.

	Without Context			With Parent			Full Context		
Category	P	R	F1	P	R	F1	P	R	F1
Neutral	90.49	89.56	90.02	90.20	90.28	90.24	90.94	88.73	89.82
AffiliationAbuse	30.30	43.90	35.86	37.80	39.41	38.59	34.03	47.30	39.58
IdentityAbuse	49.30	34.05	40.28	36.04	30.43	33.00	35.89	36.38	36.14
PersonAbuse	24.23	39.23	29.96	30.53	37.66	33.72	30.97	31.96	31.46

Table 5.3: Performance Metrics of Multi-Class Classification Models

The differentiation in model performance becomes more pronounced within abuse-related categories. The Model Without Context shows considerably higher precision in 'Identity Abuse' at 49.30% compared to 35.89% for the Model with Full Context, indicating its relative strength in minimizing false positives when context is absent. This could be due to the model's focus on explicit indicators of abuse without the potential noise introduced by additional context. Conversely, in 'Affiliation Abuse', the Model With Full Thread Context achieves the highest recall at 47.30%, suggesting it benefits from contextual cues to capture a broader range of true abusive instances. This is consistent with the notion that a full-thread context provides more cues that are indicative of abuse, even if it risks capturing more false positives, as indicated by the lower precision. The Model With One Parent in ’Person Abuse’ has the highest recall at 37.66%, indicating that a single level of context is sometimes optimal in capturing instances of personal abuse without overwhelming the model with too much information, which might be the case with full-thread context leading to a lower recall of 31.96%.

5.2.2 binary Classification Models

The binary classification models followed similar trends as the multi-class models. Without context, the model achieved a recall of 65% for ’Abusive’ speech, indicating ability to detect abusive cases despite lower precision (40%). With one parent context, the recall increased to 0.71 while precision dropped to 37%. The full thread context model saw a further recall increase to 73%, with precision remaining low at 0.36. These results suggest additional context helps improve recall in identifying abuse, but at the cost of reduced precision due to more false positives. The pattern aligns with the precision/recall trade-off observed in multi-class models.

	Without Context			With Parent			Full Context		
Category	P	R	F1	P	R	F1	P	R	F1
Neutral	0.92	0.81	0.86	0.93	0.75	0.83	0.93	0.74	0.82
Abusive Speech	0.40	0.65	0.50	0.37	0.71	0.49	0.36	0.73	0.49

Table 5.4: Binary Classification Performance Metrics by Context

5.2.3 Interpretation and Implications

These findings suggest that the performance of classifiers is affected by the extent of contextual information they incorporate. While the overall accuracy remains relatively stable, there is noticeable variation in precision and recall across different abuse categories when context is considered. For example, in multi-class models, according to the table 5.3, the model with full thread context, which takes into account the broader conversational context, tends to have higher recall, especially in categories like Affiliation Abuse (47.30%) and Identity Abuse (36.38%), compared to the model without context (43.90% and 34.05%, respectively). This increased recall can be attributed to the model’s ability to detect subtleties in the conversation that may indicate abuse, but this also leads to a higher rate of false positives, affecting precision negatively. The precision for Affiliation Abuse and Identity Abuse in the full thread context model (34.03% and 35.89%) is lower than the model without context (30.30% and 49.30%). On the other hand, the models without context, which lack this broader perspective, show higher precision in categories like Identity Abuse (49.30%). This is likely because it evaluates text in isolation, making it more effective in identifying clear-cut cases of abuse without being misled by contextual content. However, this comes at the cost of lower recall (34.05%), as the model may miss more context-dependent instances of abuse.

5.3 Conclusion

This chapter showed that there are tradeoffs when context is used to detect abusive language. Adding context by using previous comments helped BERT to increase recall in identifying abusive instances. However, it also increased false positives, reducing precision. The models that did not use context showed opposite behavior. They had higher precision, but their recall was lower. The findings highlight the challenge of maximizing both precision and recall when context is incorporated into the model.

Chapter 6

Error Analysis

6.1 Analysis of Model With Context

In conducting the error analysis for the toxic language detection model, the performance of the binary model, which incorporates the full thread context, was selected for a detailed examination. The binary model was chosen for error analysis as it provides a clear distinction between abusive and non-abusive content. By simplifying the categories, the study could better focus on the model's performance in identifying the presence or absence of toxicity, without getting distracted by the nuances of different abuse types.

A quantitative approach was adopted to sample selection. Utilizing Cochran's sample size formula, adjusted for finite populations, a representative sample of 100 errors from a total population of 1,403 that were identified misclassifications were randomly selected. The sample size was calculated to achieve a 95% confidence level with a 7% margin of error, assuming an estimated proportion of 0.8, indicative of the prevalence of errors of interest in the dataset. This statistical framework ensures that the sample is sufficiently large to be representative of the broader dataset while maintaining a manageable scope for in-depth manual review. However, as one of the limitation of this analysis, It should be noted that this random selection process did not take into account the different primary abuse categories. While this approach ensures that the sample is representative of the overall error distribution, it may not capture the nuances of errors specific to each abuse type.

The initial categories for classification were informed by "Challenges for Toxic Comment Classification: An In-Depth Error Analysis" by Van Aken et al. (2018). Comparing the error classes from Van Aken et al. (2018) with those identified in this research, several similarities and differences can be observed. some of the error classes was directly adopted from their work, such as 'Toxicity without swear words', 'Idiosyncratic and rare words', and 'Quotations or references', as these represent common challenges in toxic language detection. However, this research also introduced new error classes specific to context-aware classification, such as 'Confusion due to context', which highlights the model's difficulty in distinguishing between a comment's content and its surrounding context. Furthermore, while Van Aken et al. (2018) identified 'Sarcasm and irony' as a significant source of errors, it was found that this category is less prevalent in this analysis, possibly due to differences in the datasets or the impact of contextual information on detecting sarcasm and irony.

The error classes outlined in this research provided a structured taxonomy to categorize the model's misclassifications. A break down of these errors can be viewed in table 6.1. In the analysis phase, the frequency of occurrences within each error class was quantified and reported as a percentage of the total false positive and false negative errors.

Error Class	False Negatives	False Positives
Undersensitivity	40.00%	-
Doubtful Labels	20.00%	12.50%
Idiosyncratic and Rare Words	20.00%	6.25%
Toxicity without Swear Words	20.00%	-
Controversial Subjects	-	40.00%
Oversensitivity	-	12.50%
Confusion Due to Context	-	10.00%
Humor	-	2.50%
Usage of Swear Words	-	8.75%
Quotations or References	-	7.50%
Metaphors and comparisons	-	-%
Sarcasm and irony	-	-%

Table 6.1: Error Classes and Their Percentages

6.1.1 Error Classes of False Negatives

1 Undersensitivity

Undersensitivity 40% here refers to instances where the model fails to identify explicit or blatantly offensive language as hate speech, despite the presence of clear indicators of toxicity. While this category is a subset of false negatives, it specifically focuses on cases where the model misses unambiguous and overt expressions of hate speech. This category highlights the model's inability to recognize explicit toxicity, which is a more specific issue than general false negatives.

> **Parent Text:** "Usually people like these stick together, so you won't run into them IRL unless you're part of the group. I used to be an sjw, so I would have sjw friends and we would and say stuff like this pretty frequently online but not irl. Usually insanity like this stays on the internet. For a long time, also, I used to identify as genderfluid, but I would only say it online and not in real life because I was afraid of being made fun of. TLDR most people who write stuff like this on the internet won't admit to it IRL."
> **Text:** "People get so brave saying off-base offensive shit when there's no one present to confront them about it."
>
> **Label Info:** Abusive Speech
> **Prediction:** Neutral
> **Original Label:** `['AffiliationDirectedAbuse']`

In this example, the model potentially misinterpreted the text as a general statement rather than recognizing it as a direct attack. The phrase "brave saying off-base offensive shit" is an explicit expression of contempt, particularly targeting individuals associated with left-wing politics. Despite its blatant nature, the model's failure to correctly classify this as hate speech exemplifies the challenge of detecting direct, offensive language within a broader context. Undersensitivity constitutes a significant portion of the errors in this model, accounting for 40.00% of the cases. This high rate indicates a critical area for improvement in the model.

2 Doubtful Labels

'Doubtful Labels' 20% in hate speech detection refer to instances where the original labeling of comments is questionable or debatable, particularly when considering the context and the class definition of hate speech. 20% of these labels are in my view belongs to this category. In these cases, comments may contain elements that superficially resemble hate speech or use potentially offensive words, but upon closer examination, they do not fulfill the criteria of toxic language as per annotators' definition.

Example 1:

> **Parent Text:** “Ever think 'you know, maybe the bible was onto something when it condemned homosexuals and said don't tolerate them”
> **Text:** “Idk maybe you've got a point worth hearing, but we separate church and state in the US for good reason.”
>
> **Label Info:** Abusive Speech
> **Prediction:** Neutral
> **Original Label:** `['IdentityDirectedAbuse']`

While the parent text contains homophobic sentiments, the response is more neutral, focusing on the principle of separating church and state rather than engaging with the homophobic content. The response can be seen as an attempt to redirect the conversation toward a constitutional discussion, thereby not fulfilling the criteria of hate speech.

Example 2:

> **Parent Text:** “Look buddy, just because he said a racist word so casually does not mean he's racist.”
> **Text:** “ * ‘Context Matters!’* The context that he's willing to say racist things publicly? ”
>
> **Label Info:** Abusive Speech
> **Prediction:** Neutral
> **Original Label:** `['PersonDirectedAbuse']`

The response here challenges the downplaying of racist language in the parent text by emphasizing the misusage of context. It suggests that the willingness to use racist language publicly is indicative of racist attitudes. The text critically addresses casual racism without being toxic itself.

The subjectivity of these labels is significant for several reasons. Firstly, it acknowledges that the determination of what constitutes hate speech can vary based on individual perceptions, cultural backgrounds, and contextual understanding. This variability can lead to disagreements even among human annotators and thus, introduces a layer of complexity in the training and performance of automated detection models. Secondly, these 'Doubtful Labels' highlight the potential for confusion—not just for human reviewers but also for the models trained on such data. When a model encounters language that is ambiguously labeled or sits on the borderline of the hate speech spectrum, it may struggle to accurately categorize such content.

3 Idiosyncratic and rare words

Idiosyncratic and Rare Words 20% includes instances where the text contains unusual, less commonly known, or slang terms that the model may not recognize or may misinterpret. This category is particularly challenging because it often involves language that is not standard, is culturally or contextually specific, or may be newly coined.

> **Parent Text:** No Parent text - The following text is the first post.
> **Text:** “This heygui very traditioner but also very open. Hen scare for both Rocky and Rainy. ”
>
> **Label Info:** Abusive Speech
> **Prediction:** Neutral
> **Original Label:** `['IdentityDirectedAbuse']`

Here, the word “heygui,” as understood, is a racial slur derogatory towards Black people that originated in China. And, The reference to “Rainy” is a slang term used on online forums, particularly Reddit, to describe English-speaking Chinese girlfriends. The presence of such idiosyncratic and rare words poses a significant challenge for hate speech detection models (Zhang and Luo, 2019). These models are typically trained on datasets that may not include sufficient examples of such language, leading to a gap in understanding contextually specific or emerging terms. This category underscores the need for continuous updating and diversification of the training datasets used for hate speech detection models. Including a broader range of linguistic expressions, especially those emerging from specific cultural or internet subcultures, is essential.

4 Toxicity without Swear Words

’Toxicity without Swear Words 20% involves instances where comments contain toxic, hostile, or derogatory content but do not use explicit swear words or traditionally offensive language. This subtlety poses a significant challenge for models that predominantly rely on explicit indicators of hate speech.

> **Parent Text:** “Never could work out why Xiaojie is such a bad thing to call a girl in China.”
> **Text:** “Why is there no real word for logic in Chinese? Sometimes you don’t question things”
>
> **Label Info:** Abusive Speech
> **Prediction:** Neutral
> **Original Label:** `['IdentityDirectedAbuse']`

In this example, the response may seem innocuous at first glance. However, it subtly implies a stereotype or a derogatory view towards Chinese culture, insinuating a lack of logic or rationality. This type of toxicity is nuanced and does not rely on explicit swear words, making it harder for models to detect. The challenge in detecting ’Toxicity without Swear Words’ lies in the model’s limited capacity to grasp the subtleties of cultural contexts and the nuances of indirect expressions. For example, a comment implying a stereotype about the Chinese language and culture suggests toxicity through insinuation, not through explicit language. This subtlety of expression, where offensive sentiments are veiled under seemingly neutral language, presents a significant detection challenge.

6.1.2 Error Classes of False Positives

1 Controversial Subjects

In the 'Controversial Subjects' category, which accounts for a significant 40.00% of false positives, the model erroneously classifies discussions about sensitive or contentious topics as abusive speech. This misclassification often occurs due to the strong language or emotional intensity associated with such subjects, even when the content itself is not hateful or abusive.

Examples

> **Text:** " Dutch girl, 17, who was sexually abused at 11 and raped as a 14-year-old is legally euthanised at her home."
>
> **Label Info:** Neutral
> **Prediction:** Abusive Speech

Here, the model incorrectly flags a factual news statement as abusive speech. The serious and sensitive nature of the topic, coupled with the blunt presentation of facts about abuse and euthanasia, likely triggers the model's misclassification.

> **Text:** " Soviets commited genocide against Germans, Ukrainians, Crimean Tartars plus Chechens, Koreans, Poles. All specifically targeted for their ethnicity.."
>
> **Label Info:** Neutral
> **Prediction:** Abusive Speech

Despite being a historical discussion, the model wrongly interprets this text as abusive, possibly due to the mention of genocide and ethnic targeting, which are controversial and loaded topics.

> **Text:** "I agree with the sentiment of your post but I'm not sure I agree euthanasia is part of the larger conspiracy. Mass-migration, the media telling 'white' people to have less babies.. yeah, definitely subversive and calculating... But part of the larger plan.. I just don't see it."
>
> **Label Info:** Neutral
> **Prediction:** Abusive Speech

This example involves a discussion on several contentious issues, including euthanasia and demographic changes. The model flags it as abusive, likely due to the combination of sensitive topics and the speculative nature of the statements.

While controversial subjects naturally involve emotional intensity and strong language, these topics are also often discussed alongside hate speech and toxic rhetoric. As a result, it is plausible that the model has learned an association between such subjects and abusive content during training. This learned association likely contributes to the model misclassifying discussions around controversial issues as abusive speech, even when the language itself is not hateful. The challenge is amplified when dealing with subjects that evoke strong sentiments or involve complex historical and ethical nuances, which can resemble patterns of hate speech the model was exposed to during training.

The high percentage of false positives in this category suggests a tendency of the model to be extra caution, flagging potentially contentious content as abusive to avoid missing genuine cases of hate speech. However, this approach leads to a significant number of misclassifications, impeding the discussion of important but sensitive topics.

2 Doubtful Labels

As discussed in the discussion of false negatives, 'doubtful labels' are also presented in False positives. There are 12.50% samples where a post has been marked as neutral, but in the subjective view of the author of this thesis, they should have been marked as harmful speech.

> **Parent Text:** " I love the absolute right wing takeover of this subreddit, it gives me hope."
> **Text:** " This is barely right wing stfu."
>
> **Label Info:** Neutral
> **Prediction:** Abusive Speech

In this example, the model identifies "stfu" (shorthand for "shut the f*** up") as harmful speech and a personal attack, leading to its classification as 'Abusive Speech'. However, the annotators marked this as 'Neutral'.

3 Oversensitivity

'Oversensitivity' accounts for 12.50% of false positives in the classification. While oversensitivity errors result in false positives, this category emphasizes the model's tendency to overreact to specific language cues, even when they are used in a non-toxic context. This is different from general false positives, which can arise due to various reasons. Oversensitivity specifically points to the model's excessive reliance on certain words or phrases as indicators of toxicity, resulting in misclassifications. This is in stark contrast to the previously discussed 'Undersensitivity' category, which constituted a much larger proportion (40.00%) of errors. Oversensitivity highlights the model's propensity to overreact to content that lacks clear indicators of offensiveness.

Example

> **Parent Text:** "What kin type are you?"
> **Text:** "Type 1 poisonous fart-kin. I really need to start eating healthily."
>
> **Label Info:** Neutral
> **Prediction:** Abusive Speech

In this case, the speaker's self-deprecating comment about their eating habits is flagged as 'Abusive Speech', despite no apparent malice or hate speech.

This example illustrates the model's oversensitivity in categorizing content. The phrase "Type 1 poisonous fart-kin" is self-referential, lacking any abusive or hateful intent. The model's misclassification here likely stems from its inability to understand the playful tone of the comment, as well as the lack of contextual cues that would indicate toxicity.

The lower incidence of oversensitivity compared to undersensitivity suggests that the model is more prone to missing explicit hate speech than it is to over-flagging neutral content. However, the presence of oversensitivity errors is still significant, as it can lead to the unwarranted censorship of harmless expressions and impede free speech.

4 Confusion Due to Context

Confusion Due to Context, representing 10.00% of false positives, occurs when the model misclassifies a statement as abusive speech due to being influenced by the surrounding context. This category highlights the model's challenge in accurately separating the text from the parent text within its conversational thread.

Example:

> **Parent Text1:** "It's becoming a crime to discuss/question events"
> **Parent Text2:** "This post is most likely inspired by the Christchurch shootings, so let's make things clear. Those videos were removed because they served no purpose other than to glorify what that sick bastard did. (ignoring the fact that it's been illegal to share shit like that in New Zealand since 2002) (if you're interested, the law banning possession of mass shootings also bans child porn, rape videos and other fucked up shit from the internet/media in general)."
>
> **Text:** "You're acting like there's no difference between porn and this."
>
> **Label Info:** Neutral
> **Prediction:** Abusive Speech

The model incorrectly classifies the text as 'Abusive Speech'. The model likely arrives at this incorrect classification due to the influence of the preceding 'parent text2', which includes strong and explicit language ("sick bastard", "shit like that") to describe the person and the event of a mass shooting. The key issue here is that the model seems to be overly influenced by the context provided by the preceding messages in the conversation thread. While context is crucial for understanding the nuances of a conversation, in this case, the model fails to correctly separate the sentiment of Speaker1's text from the emotional intensity of the preceding messages. This is also supported by the fact that the model without parent text does not classify these comment as abusive.

5 Humor

Only 2.50% of false positive samples in our dataset were due to humor misclassification, indicating that this is not a predominant issue in this dataset.

Example

> **Parent Text1:** " Sometimes I wonder how many neighborhoods I could buy in Detroit with my life savings."
> **Parent Text2:** " 150 acres of urban blight."
> **Text:** " Bulldoze it all down and sell it to the Chinese."
>
> **Label Info:** Neutral
> **Prediction:** Abusive Speech

The joke here lies in the hyperbolic and absurd nature of the proposed solution. It's not a serious proposal but rather a satirical take on urban redevelopment; potentially also poking fun at the common trope of foreign investment in real estate. As illustrated by the example, the primary challenge in detecting humor lies in the subtlety of its expression. Humor can be highly context-dependent, relying on the audience's understanding of underlying assumptions, shared knowledge, or the absurdity of the situation. For a machine learning model, decoding these nuanced cues is complex, as it requires an understanding beyond the literal meaning of words (Winters and Delobelle, 2020; Smadu et al., 2021).

6 Usage of Swear Words'

The category 'Usage of Swear Words' in false positives, accounting for 8.75% of the cases in our dataset, highlights a common challenge in NLP models: the overemphasis on swear words as indicators of toxicity. This issue aligns with findings from Van Aken et al. (2018) where they noted a similar pattern in other datasets.

Example

> **Parent Text1:** " Because if Matthew is simply killing to kill and not for political or religious reasons, he's not a terrorist. If Ahmed kills for political or religious reasons, he's a terrorist. Like I said, the motive is directly related to the definition of terrorism."
> **Parent Text2:** "So the terrorism is only for muslims? If Matthew kills people in a mosque, he's simply just sick?Not an act of terror? "
> **Text:** "Jesus fucking Christ. Are you a troll or can you legitimately not read?...It has ZERO to do with race or specific religion they chose to follow."
>
> **Label Info:** Neutral
> **Prediction:** Abusive Speech

Despite the use of strong language ("Jesus fucking Christ, Are you a troll "), the speaker express frustration or emphasize a point in a heated discussion about terrorism and religion.

This example underscores the model's tendency to sometimes misinterpret the presence of swear words as a sign of abusive speech. However, in many conversational contexts, such language can be used for emphasis or as part of colloquial speech, rather than to convey toxicity or abuse. The challenge lies in the model's limited ability to interpret the contextual use of swear words. Often, swear words are employed for various reasons—emphasis, frustration, or even camaraderie—which are not inherently toxic. The model's inability to differentiate these nuances leads to a higher rate of false positives in texts containing swear words.

7 Quotations or References

The 'Quotations or References' category, accounting for 7.50% of false positives, involves instances where the model incorrectly identifies statements as hate speech when they are merely quoting or referencing potentially sensitive content in response to a previous post. This error type occurs when the model fails to recognize the context in which certain phrases are used.

Example

> **Parent Text1:** "Why are there only two genders?"
>
> **Parent Text2:** "Because evolution has not paved the way for more than two neurological pathways associated with sex and behavior.Edit: To be clear, I am talking solely about gender. There are 2 genders, the neurological constructs linking sex and behavior."
>
> **Text:** "I don't understand what you mean by this, but the phrasing seems to imply that gay people cannot exist."
>
> **Label Info:** Neutral
> **Prediction:** Abusive Speech

In this instance, the model misinterprets a part of the conversation, specifically the phrase "gay people cannot exist", as abusive speech. This misunderstanding likely stems from the model's inability to discern that the speaker is summarizing or questioning the previous speaker's viewpoint rather than expressing their own prejudiced opinion. This case illustrates the challenge faced by the model in differentiating between a user endorsing a sensitive or controversial opinion and a user simply referring to or questioning such an opinion. The misclassification here indicates a lack of contextual understanding by the model, particularly in terms of recognizing the intent and stance of the speaker in a multi-turn conversation.

8 Idiosyncratic and Rare Words

Similar to the 'Idiosyncratic and Rare Words' category discussed in the context of false negatives, this category also plays a role in false positives, accounting for 6.25% of such cases. This category involves instances where the model misinterprets or is unfamiliar with unusual, less commonly known, or slang terms.

Example

> **Parent Text:** "Those !!!are!!! diffferent (national emergency vs executive order). The more plausible Smuggie would be criticizing this 3rd national emergency of Trump whilst justifying previous Presidents amassing even more (Bush 12, Obama 13)?"
>
> **Text:** "You're reading too deep into a smuggie."
>
> **Label Info:** Neutral
>
> **Prediction:** Abusive Speech

In this case, the term "smuggie" (referring to illustrations featuring caricatures of people with extreme political viewpoints) is not recognized by the model, possibly leading it to interpret the word as derogatory slang.

6.2 Comparative Analysis of Models with and without Context

Following the initial error analysis, a comparative study was conducted to gain a deeper understanding of how the model's performance differs when context is taken into account versus when it is not. To quantify the performance difference between the models with and without context, first, the predictions from both models (with and without context) were compared to identify instances where the two models disagreed. From this filtered dataset of disagreements (591 instances in total), a random sample of 100 instances was selected. This random sampling ensures that the analysis is based on a representative subset of the instances where the models had conflicting predictions. When examining the agreement between the model predictions and the ground truth labels, the model without context agreed with the ground truth in 68% of the cases, while the model with context agreed with the ground truth in 32% of the cases.

While the model without context was better represented in this random sampling, a comparative error analysis showed that there were several instances where the model with context excelled. In particular, the model with context demonstrated improved performance in cases where the context was crucial for accurately interpreting the meaning and intent behind a comment.

Since the previous sections cover the errors caused by disagreement between the model with context and the annotation, this section of the analysis will focus on analyzing cases where context makes a difference.

1 Context Matters

In 17% of the cases where the model with context agreed with the ground truth (which constitutes 32% of the total sample), the analysis showed that context played a crucial role in the correct classification.

For example, consider the following interaction:

Parent Text: “Bush isn’t anybody’s friend.”

Text: “but he smokes WEED, DUDE.”

Label Info: Neutral

Prediction model without context: Abusive Speech

Prediction model with context: Neutral

In this case, the model without context misclassified the comment as abusive speech, likely due to the presence of the word “WEED” in all caps. However, when considering the context provided by the parent text, it becomes clear that the comment is a humorous or sarcastic response to the statement about Bush. The model with context correctly identifies the comment as neutral, demonstrating the importance of contextual information in accurately interpreting the intent behind the comment.

Another example that highlights the significance of context is:

Parent Text 1: “I’ve been told everything is politics at this point. So I guess the don’t talk about ”politics or religion“ rule means just shut up and never talk about anything.”

Parent Text 2: “This one is starting to understand.”

Text: “We should kill him before things get out of hand.”

Label Info: Abusive Speech/PersonDirectedAbuse

Prediction model without context: Neutral

Prediction model with context: Abusive Speech

In this interaction, the model without context misclassifies the comment as neutral, failing to recognize the abusive nature of the statement "We should kill him." However, when considering the context provided by the parent texts, the model with context correctly identifies the comment as abusive speech. The context helps to reveal the threatening and harmful intent behind the comment.

2 context not needed

In 15% of the cases where the model with context agreed with the ground truth (which were 32% of the total sample), the model with context classified the comment correctly, while the model without context did not. However, the context of previous comments was not necessary for this classification. It appears that the model trained with context was simply better able to discern whether a comment is abusive or not, even without relying on the specific context.

For instance, consider the following example:

> **Parent Text:** "what got them into the fight?"
>
> **Text:** "Alcohol and testosterone, by the looks of it."
>
> **Label Info:** Neutral
>
> **Prediction model without context:** Abusive Speech
>
> **Prediction model with context:** Neutral

In this case, the model without context misclassified the comment as abusive speech, possibly due to the mention of "alcohol" and "testosterone." However, the model with context correctly identified the comment as neutral, even though the parent text does not provide any essential context for the classification. This suggests that the model trained with context may have developed a better understanding of what constitutes abusive speech, independent of the specific context in which the comment appears.

These samples highlight that while context can be crucial in some cases, there are also instances where the model trained with context performs better in classifying comments, even when the specific context is not necessary for the classification. This points to the potential benefits of training models with contextual information, as it may improve their overall ability to distinguish between abusive and non-abusive speech.

6.3 Conclusion

The error analysis presented several key findings regarding the performance of the toxic language detection model that incorporates full thread context. In false negatives, undersensitivity (40%), where the model failed to detect explicit offensive language, was the category that amounted for the majority of the errors. In false positives, the top category was controversial subjects (40%), where the model wrongfully associated sensitive topics, which are often associated with hate speech, as abusive content.

The error analysis also revealed that while the "Confusion Due to Context" category, which highlighted interesting instances when the model misclassified a statement as abusive speech due to being influenced by the surrounding context in the conversational thread, accounted for 10% of false positives, and it was not a highly frequent error.

In the comparative analysis between the model with full context and the model without context, the analysis showed that context played a crucial role in accurate classification in some instances (17% of time) where the model with context was correctly able to identify the nature of the comment. The context that was provided through previous comments enabled the model to disambiguate intent and appropriately interpret statements that may have seemed abusive or neutral in isolation.

Finally, the comparative study showed that in a subset of cases, the context model performed better without relying on the specific context. This finding suggests that while context can be advantageous, it may also introduce noise or confounding factors in certain situations. All in all, while providing context can aid classification, it also presents its own set of challenges that need to be addressed.

Chapter 7

Conclusion

This research has shown that the inclusion of context does not substantially enhance the performance of BERT models in detecting toxic language within the CAD. One of the key insights comes from the annotation analysis in Chapter 3. One of the characteristics of this dataset is that the annotators relied on the context of previous comments only 6% of the time. In the remaining 94% of cases, the annotators were able to accurately classify the comments solely based on the content of the individual comment itself. This finding suggests that most comments in CAD can be understood and classified correctly without the need for additional context. This low dependency on context indicates that the majority of abusive content in this dataset is explicit enough to be detected without the previous comments. As a result, this dataset may not be best suited to train models to learn from context. The error analysis further supports this conclusion, as 'Confusion due to Context' accounted for only 10% of false positives, a relatively small proportion compared to other error categories.

The error analysis provides further insights into the model's performance. One of the key findings from error analysis was "Undersensitivity" to toxic language, where the model failed to identify explicit or blatantly offensive language as hate speech despite clear indicators of toxicity. This accounted for 40% of false negatives, highlighting a substantial gap in the model's ability to recognize direct and overt expressions of hate speech. Another critical category is the misinterpretation of idiosyncratic and rare words, which contributed to 20% of false negatives and 6.25% of false positives. On the other hand, "Controversial Subjects" accounted for 40% of false positives, indicating that the model frequently misclassified discussions on sensitive or controversial topics as abusive. Similarly, "Oversensitivity" to perceived toxicity was observed in another 12% of false positives, where the model incorrectly identified non-abusive comments as toxic language due to wrongly perceived indicators of toxicity. The relatively low percentage of errors related to confusion due to context, combined with the findings from the annotation analysis, indicates that the model's performance issues are more related to other factors, such as the inherent difficulty of the language or subject matter, rather than the presence or absence of context.

However, it was also noteworthy that the comparative error analysis between the model with and without context showed that context could still play an important role in certain cases. While the annotation analysis and the overall error breakdown suggest that context is not a predominant factor for this dataset, there were instances where considering the surrounding conversational context was crucial for accurate classification. In approximately 17% of the cases where the model with context agreed with the ground truth, the analysis revealed that context played a vital role in enabling the correct prediction. The comparative analysis demonstrated scenarios where the isolated comment could be misinterpreted without the contextual information provided by the previous statements. In this small subset, factors such as sarcasm, humor, and implicit threats or harmful intent were better captured when the model had access to the context that was provided by previous posts.

One possible path for future research is to develop better methods for automatically identifying when context is needed. For instance, Vidgen et al. (2021) have identified in their dataset instances where the annotators relied on context to determine the nature of the text (whether it is toxic or not). By utilizing these existing annotations, future studies could selectively incorporate context only in the cases where annotators have indicated that context was needed. This targeted approach would allow for an examination of whether providing contextual information improves BERT's classification performance specifically in scenarios where human annotators deemed context as crucial, without indiscriminately adding context to all samples. By evaluating the model's performance on these annotated instances, both with and without the additional context, researchers could gain valuable insights into the potential benefits of selectively incorporating context when it is deemed necessary according to human judgment. This approach leverages the existing annotations and human judgments available in datasets like the one used by Vidgen et al. (2021), which could serve as a valuable starting point for identifying instances where context is likely to be beneficial. By isolating these cases and evaluating the impact of context on model performance, future research could potentially uncover more nuanced findings and develop more effective strategies for leveraging contextual information in hate speech detection and related natural language processing tasks.

Another potential direction is to explore alternative approaches to incorporate context more effectively. The current method used in this research involved concatenating the main text and parent text before feeding it into the BERT model. However, this may not be the optimal approach. Future research should investigate the use of separate BERT models for the main text and parent text, allowing each model to capture the distinct features and representations more effectively. Weighted concatenation techniques, where the representations from the main text and parent text models are assigned different weights before concatenation, could also be explored. This would emphasize the importance of the main text while still considering the contextual information. Additionally, researchers could experiment with attention mechanisms or other architectures that can dynamically learn to focus on the relevant parts of the context based on the main text.

Overall, while the findings suggest that context is not a predominant factor for this particular dataset, there are promising avenues for future research to leverage contextual information more effectively, potentially leading to improved performance in hate speech detection and other natural language processing tasks.

Bibliography

V. Agarwal, A. P. Young, S. Joglekar, and N. Sastry. A graph-based context-aware model to understand online conversations. *arXiv preprint arXiv:2211.09207*, 2022.

A. Baheti, M. Sap, A. Ritter, and M. Riedl. Just say no: Analyzing the stance of neural dialogue generation in offensive contexts. *arXiv preprint arXiv:2108.11830*, 2021.

D. Cortiz and A. Zubiaga. Ethical and technical challenges of ai in tackling hate speech. *The International Review of Information Ethics*, 29, 2020.

T. Davidson, D. Warmsley, M. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515, 2017.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

P. Fortuna and S. Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30, 2018.

L. Gao and R. Huang. Detecting online hate speech using context aware models. *arXiv preprint arXiv:1710.07395*, 2017.

J. A. García-Díaz, S. M. Jiménez-Zafra, M. A. García-Cumbreras, and R. Valencia-García. Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers. *Complex & Intelligent Systems*, pages 1–22, 2022.

J. B. Jacobs and K. Potter. *Hate crimes: Criminal law & identity politics.* Oxford University Press on Demand, 1998.

M. Karan and J. Šnajder. Preemptive toxic language detection in wikipedia comments using thread-level context. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 129–134, 2019.

Y. Kishimoto, Y. Murawaki, and S. Kurohashi. Adapting bert to implicit discourse relation classification with a focus on discourse connectives. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1152–1158, 2020.

J. S. Lemmens, M. Simon, and S. R. Sumter. Fear and loathing in vr: the emotional and physiological effects of immersive games. *Virtual Reality*, 26(1):223–234, 2022.

I. Markov and W. Daelemans. The role of context in detecting the target of hate speech. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 37–42, 2022.

S. Menini, A. P. Aprosio, and S. Tonelli. Abuse is contextual, what about nlp? the role of context in abusive language annotation and detection. *arXiv preprint arXiv:2103.14916*, 2021.

A. Merchant, E. Rahimtoroghi, E. Pavlick, and I. Tenney. What happens to bert embeddings during fine-tuning? *arXiv preprint arXiv:2004.14448*, 2020.

I. Mollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumakas. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, 8(6):4663–4678, 2022.

E. Saquete, D. Tomás, P. Moreda, P. Martínez-Barco, and M. Palomar. Fighting post-truth using natural language processing: A review and open challenges. *Expert systems with applications*, 141:112943, 2020.

R.-A. Smadu, D.-C. Cercel, and M. Dascalu. Upb at semeval-2021 task 7: Adversarial multi-task learning for detecting and rating humor and offense. In *SemEval@ ACL/IJCNLP*, pages 1160–1168, 2021.

B. Van Aken, J. Risch, R. Krestel, and A. Löser. Challenges for toxic comment classification: An in-depth error analysis. *arXiv preprint arXiv:1809.07572*, 2018.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

M. Verbrugge. *The BERT Ranking Paradigm: Training Strategies Evaluated.* PhD thesis, Radboud University Nijmegen, 2021.

B. Vidgen, H. Margetts, and A. Harris. How much online abuse is there. *Alan Turing Institute*, 11, 2019.

B. Vidgen, D. Nguyen, H. Margetts, P. Rossini, and R. Tromble. Introducing cad: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, 2021.

S. Walker. *Hate speech: The history of an American controversy.* U of Nebraska Press, 1994.

T. Winters and P. Delobelle. Dutch humor detection by generating negative examples. *arXiv preprint arXiv:2010.13652*, 2020.

E. Wulczyn, N. Thain, and L. Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399, 2017.

M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*, 2019.

Z. Zhang and L. Luo. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10(5):925–945, 2019.