Master Thesis

Exploring Implicit Abusive Speech Detection: A Comprehensive Analysis of Fine-Tuning BERT and Prompting Qwen2.5

K.D. Gerritsen

Supervisor	E. Maks
2nd reader	A. Fokkens

a thesis submitted in fulfillment of the requirements for the degree of

MA Linguistics
(Text Mining)

Vrije Universiteit Amsterdam

Computational Linguistics and Text-Mining Lab
Department of Language and Communication
Faculty of Humanities

Date	June 27, 2025
Student number	2723340
Word count	16,473

Abstract

Online content has become a significant part of our daily lives. This development made detecting abusive speech important for the overall well-being of society. One of the challenges in detecting abusive speech involves the nuanced and implicit ways in which it can be expressed. Since the early 2010s, researchers have made several attempts to tackle the detection of abusive forms of speech, with recent work showing promising results for transformer-based models and generative large language models (LLMs).

In this research, I aim to contribute to the detection of abusive speech by answering the question of whether prompt engineering offers advantages over a fine-tuned BERT model, particularly in identifying implicit cases of abusive speech. I conduct two main experiments — fine-tuning BERT-based models and prompting Qwen2.5 — across both binary (abusive vs. not abusive) and ternary (explicit abuse, implicit abuse, not abusive) classification tasks, and evaluate the performances on the AbuseEval test set. Finally, I conduct a thorough error analysis to examine how errors, and in particular mistakes in implicit abusive speech, affect the model's results.

The results show that fine-tuning still delivered better overall performance, achieving a macro-averaged F1-score of 0.60, with 0.29 for implicit cases. The best-performing prompting strategy combined Chain-of-Thought (CoT) with considering targetness, reaching a macro-averaged F1-score of 0.52, with 0.25 in the implicit class. The error analysis revealed that helping the model understand the boundary between explicit and implicit abuse, and implicit and non-abusive, through improving understanding of the target and the context within tweets, is key in reducing misclassification in abusive speech, particularly in implicit cases.

Declaration of Authorship

I, author, declare that this thesis, titled 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: 27-06-2025

Signed:

Acknowledgments

List of Figures

List of Tables

Contents

Chapter 1

Introduction

With the rise of ever-growing online media, 24/7 accessibility to online content has become a significant part of our lives. However, this expansion has also brought significant challenges, including the growth of abusive speech with a worldwide reach. Abusive speech can be harmful to others and contribute to causing serious consequences for people of all demographics (ElSherief et al., 2021). Therefore, successfully addressing it is important for the overall well-being of society. Since the early 2010s, researchers have made several attempts to tackle the detection of abusive forms of speech, with recent work showing promising results for transformer-based models and generative large language models (LLMs) (Caselli et al., 2021; OpenAI, 2024; Roy et al., 2023; Huang et al., 2023; Ziems et al., 2024).

Despite these growing efforts, the definition of abusive speech remains debated in NLP research, as it varies across cultures, legal frameworks, and academic disciplines, lacking a universal consensus (Caselli et al., 2020). In this context, I adopt the definition of abusive speech as "any strongly impolite, rude, or hurtful language using profanity that can express debasement of someone or something, or convey intense emotion" (Founta et al., 2018). This functions as an umbrella term, with hate speech considered a subset of abusive speech. While offensive language may be abusive, it is not necessarily so.

One of the challenges in detecting abusive speech involves the nuanced and implicit ways which it can be expressed (Ocampo et al., 2023). These forms of abusive speech can often bypass surface-level detection and even advanced models if they have not been trained on subtle forms (ElSherief et al., 2021). The complexity of detection increases the risk that individuals and groups may cause harm while avoiding accountability (ElSherief et al., 2021). The relevance of research on this matter is therefore highly significant for improving the fairness and robustness of abusive language detection in general.

Over the last few years, several researchers have focused more on the distinction between explicit and implicit abuse to improve the overall detection of abusive speech (Caselli et al., 2020; ElSherief et al., 2021; Vidgen et al., 2021; Mathew et al., 2021; Hartvigsen et al., 2022). Despite these efforts, the task remains challenging, and there is still room for improvement in building models that can decipher forms such as coded language, contextualized sarcasm, and figurative language. In this research, I aim to contribute to the detection of abusive speech by answering the question of whether prompt engineering offers advantages over a fine-tuned BERT model, particularly in identifying implicit cases of abusive speech.

1.1 Approach

This research takes a twofold approach to explore this question. I conduct two main experiments — fine-tuning BERT-based models and prompting a generative model — across both binary (*abusive* vs. *not abusive*) and ternary (*explicit abuse*, *implicit abuse*, *not abusive*) classification tasks, and evaluate the performances on the AbuseEval test set (Caselli et al., 2020).[1]

The first part investigates the performance of fine-tuned BERT and HateBERT models (Devlin et al., 2019; Caselli et al., 2021). I run four experimental setups per classification scheme to assess the influence of model choice and training data on detecting abusive speech, with a focus on implicit cases. The best-performing configuration serves as the baseline for the prompting experiments.

The second part explores which prompting strategies are most effective in improving the detection of abusive speech using a generative model. I prompt Qwen2.5 using a three-step setup — including four-shot configurations, five different prompting strategies, and testing different temperature values — to get an answer to this question (Yang et al., 2025).

Finally, I conduct a thorough error analysis to examine how errors, and in particular mistakes in implicit abusive speech, affect the model's results. This includes reviewing the overall classification report, manually analyzing error patterns in the main areas of confusion, and running three automated analyses on subsets expected to challenge the model. With this, I aim to get a grasp on what is happening inside the model and determine why certain cases of implicit speech are incorrectly predicted.

1.2 Outline

This paper is structured as follows. Chapter 2 discusses key terminology and provides an overview of related work, focusing on the distinction between explicit and implicit abuse, as well as the development in automated abusive speech detection methods. Chapter 3 presents the datasets used, including label counts and explanations of AbuseEval and the Implicit Hate Corpus (IHC) (ElSherief et al., 2021). Chapter 4 outlines the methodology employed for both the fine-tuning and prompting experiments, including a detailed summary of prompt design. Chapter 5 reports the results of both sets of experiments, while Chapter 6 offers an extended error analysis of the model output, focusing on the classification report and providing manual and automated analyses of subsets of errors. Chapter 7 discusses the findings in relation to the research questions and addresses the limitations of the study. Finally, Chapter 8 concludes the paper by providing an overview of the findings and offering final reflections and suggestions for future work.

[1] The full implementation, including code for data, model training, prompting, and evaluation, is available at: `https://github.com/kimdolly27/exploring-abusive-speech`.

Chapter 2

Terminology and Related Work

This chapter provides an overview of the key terminology and prior research on the detection of abusive speech. Section 2.1 defines abusive speech and highlights the distinction between explicit and implicit forms. Section 2.2 summarizes previous related work on automated abusive speech detection, covering traditional and recent machine learning approaches.

2.1 Terminology

2.1.1 Defining abusive speech

Determining what includes abusive speech is inherently challenging. Interpretations vary between cultures, legal systems, and academia, leaving its definition in NLP research an ongoing debate without a universal consensus (Caselli et al., 2020). Moreover, abusive speech is frequently used interchangeably with several other terms such as hate speech, offensive language, toxic language, or cyberbullying (Fortuna and Nunes, 2018). The lack of consensus has resulted in contradictory guidelines and annotations; therefore, interpretation or comparison should be approached with caution (Caselli et al., 2020).

Among these overlapping terms, **hate speech** is the most universally recognized and formally defined. In the United States, hate speech is protected under the free speech provisions of the First Amendment, although it has been extensively debated in the legal sphere. In many other countries, including the United Kingdom, Canada, and the Netherlands, there are laws that prohibit hate speech, typically defined as speech that targets minority groups in a way that could promote violence or social disorder (Davidson et al., 2017). These differences across countries make detecting hate speech a context-dependent and subjective task; what is labeled as hate speech in one legal or cultural setting may not be considered as such in another.

The research by Davidson et al. (2017) is one of the first benchmark papers on the distinction of terms in NLP. From then on, researchers started recognizing different forms that require more nuanced detection. The authors identify the distinction between offensive language and hate speech as a central challenge in the automatic detection of hate speech. Although there is no formal definition in NLP, there is a consensus that hate speech is language that targets disadvantaged social groups in a manner that is potentially harmful to them (Davidson et al., 2017). Following this, Davidson et al. (2017) defines hate speech as language that is used to express hatred towards a targeted group or is intended to be derogatory, humiliating, or insulting the members of the group. Other research states that hate speech can either be directed towards a specific individual or entity, or can be used towards a generalized other (Waseem et al., 2017). Targetedness is essential for the definition of hateful speech and has been captured through fine-grained labels in several studies (Davidson et al., 2017; Waseem et al., 2017; Zampieri et al., 2019; Caselli et al., 2020).

Although all hate speech is offensive, not all **offensive speech** is hate speech. The difference between the two often relies on subtle linguistic distinctions, many of which can be ambiguous (Davidson et al., 2017). Offensive speech may be profane or socially inappropriate, but does not necessarily express targeted hate toward a group or individual. The definition of hate speech provided above does not cover all instances of offensive language, as people often use highly offensive terms to certain groups in a qualitatively different way (Davidson et al., 2017). For example, slurs or profane terms such as *'sl*t'* or *'n*gga'* may be used among peers or in music lyrics within specific communities. These terms can be considered offensive but not necessarily hateful. In any case, it is important to consider that if the distinction between hate and offensive speech is not made, models trained to detect hate speech may learn to detect offensive language, which can introduce bias in both annotation practices and automated detection systems (Davidson et al., 2017).

Another term that has gained popularity in NLP over the last few years, and whose associated task this paper will examine in more depth, is **abusive speech**. When it comes to hate speech, offensive speech, and abusive speech, the distinguishing factor is their level of specificity (Caselli et al., 2021). While abusive speech is sometimes used interchangeably with offensive language, abusive language is more often defined as an umbrella term that includes hate speech and offensive language (Waseem et al., 2017). A statistical analysis by Founta et al. (2018) showed that categories such as abusive and offensive tend to be significantly correlated, highly coexisting, and very similar. Based on these findings, the authors ultimately argue to merge the overlapping categories into the label abusive as it covers the widest range of these concepts (Caselli et al., 2020). They state the following definition for abusive speech: “any strongly impolite, rude, or hurtful language using profanity, that can show a debasement of someone or something, or show intense emotion” (Founta et al., 2018).

In my research, I adopt this definition when referring to abusive speech, where hate speech is seen as a subset of abusive speech. While offensive speech may be abusive, it is not necessarily so. Given that not all previous studies make a clear distinction between terms and that there is significant overlap in meaning, this study also thoughtfully incorporates earlier research specifically focused on detecting what is considered hate speech.

2.1.2 Explicit vs. implicit

One of the main challenges in detecting abusive speech is the nuanced ways in which abuse can be expressed. While explicit abuse is relatively straightforward to detect by recognizing specific abusive words, the real challenge for automated systems is to identify language containing linguistically subtle and implicit forms (Ocampo et al., 2023). The first approaches to compiling datasets for hate speech detection did not explicitly focus on the relevance of this distinction. However, in the last five years, the topic has gained more awareness (Caselli et al., 2020; ElSherief et al., 2021; Vidgen et al., 2021; Mathew et al., 2021; Hartvigsen et al., 2022).

Explicit and implicit abusive language have different linguistic features that are important to consider during the development of detection systems. In the case of **explicit** abusive speech, the language is unambiguous in its potential to be abusive: the literal definition of the words used according to the dictionary is abusive, such as language containing racist or homophobic slurs or profanity (Ocampo et al., 2023). It is more informal, angrier, and often explicitly attacks the target, making use of second-person pronouns or specific names, with fewer analytic words and more words suggesting authority and influence (Zampieri et al., 2019; ElSherief et al., 2018). The sentences are often short, fragmented, and grammatically simple, and imperative forms and aggressive syntactic structures are prevalent (Waseem and Hovy, 2016; Davidson et al., 2017). The following example illustrates explicit abuse through profanity and direct insult, targeting a person in the second person, with no ambiguity or contextual interpretation and use of imperative form.

> "You're nothing but a worthless f**king idiot, go back to your own country!" [1]

Explicit slurs and profanity are overall absent in **implicit** abusive speech, which relies instead on more subtle and indirect forms of abuse. In an extensive study of hate speech datasets, Ocampo et al. (2023) identified 18 typical properties of implicitness based on linguistic characteristics. Among the most significant are irony, sarcasm, black humor, metaphor, exaggeration, rhetorical questions, sentiment, inference, lack of context, and absence of extralinguistic knowledge, often appearing in combination (Ocampo et al., 2023). The following example illustrates abusive language in an implicit form through sarcasm and coded language. It implies racial or social bias without using overtly offensive or profane words.

> "Oh great, another one of them got the job. Must be all that 'diversity' working its magic again."[1]

This type of language can often bypass surface-level detection and even advanced models if they have not been trained on subtle forms. The complexity in detecting increases the risk that individuals and groups cause harm while avoiding accountability (ElSherief et al., 2021). The relevance of research on this matter is therefore highly significant for improving the fairness and robustness of abusive language detection in general.

Over the last few years, we have seen a shift in the development of annotated datasets towards incorporating this distinction. The first annotation studies that addressed this are AbuseEval v1.0 by Caselli et al. (2020) and Implicit Hate Corpus (IHC) by ElSherief et al. (2021). Both studies proposed annotation guidelines that distinguish between explicit and implicit abuse. Following this, Vidgen et al. (2021) introduced an "animosity" label in their Contextual Abuse Dataset, similar to what is considered implicit. More recently, HateXplain by Mathew et al. (2021) used human-annotated rationales, providing not only labels for each post but also highlighting the parts of the text on which the labeling decision is based. Another attempt was the ToxiGen dataset by Hartvigsen et al. (2022), which uses machine-generated statements about minority groups to better train models on more subtle forms of hate speech. These

[1] Generated by ChatGPT on May 19, 2025. Full prompts can be found in Appendix A.

studies show a growing interest in NLP to capture more complex and implicit forms of abusive speech.

Despite these efforts, detecting implicit abuse remains challenging. There is still room for improvement in exploring deciphering models for coded language, contextualized sarcasm and figurative language detection, and bias mitigation in abusive speech detection systems and their connection to the data set (ElSherief et al., 2021; Caselli et al., 2020). These challenges underscore the need for further research focused on improving the detection of subtle and implicit forms for the overall quality of abusive speech detection.

2.2 Automated detection

Research on detecting abusive speech has been driven by the rapid growth of online platforms such as Facebook, Twitter, Reddit, and Instagram in the early 2010s. As these platforms expanded, so did public awareness of the harm caused by abusive language and the need for automated moderation. Since the early 2010s, abusive speech detection systems have evolved from rule-based approaches to large language models (LLMs) that are increasingly capable of detecting the nuance and complexity of language.

2.2.1 From Rule-Based to Deep Learning

The first attempts at abusive speech detection were rule-based and lexicon-based systems, using a hate speech lexicon combined with syntactic pattern matching (Warner and Hirschberg, 2012). However, because of their reliance on keywords and surface-level patterns, these methods were very limited in their performance.

Supervised learning machine learning techniques significantly improved some of the original limits encountered in rule-based systems. The use of methods such as logistic regression marked an important shift away from keyword matching toward techniques that could better handle the informal language of social media (Waseem and Hovy, 2016; Davidson et al., 2017). By using hand-crafted features including character n-grams, word n-grams, and part-of-speech tags, these models were able to capture linguistic patterns often associated with abusive speech more effectively. Although these models show improvements, they still rely on surface-level features. As a result, models trained on one dataset often fail to generalize well to others, which increases the risk of overfitting to the training data.

A comparative study of classical machine learning and deep learning models, such as CNNs and LSTMs, demonstrated a significant performance advantage for the deep learning approaches. During the task of classifying a tweet as racist, sexist, or neither, the best results were obtained using an LSTM model with randomly initialized embeddings (Badjatiya et al., 2017). Deep learning methods learn linguistic patterns directly from the data, which gives them an advantage over models that rely on handcrafted features. However, these approaches remain limited in their ability to detect less explicit or context-dependent forms of abusive speech.

2.2.2 The transformer era

A shift in NLP was marked by the introduction of BERT (Bidirectional Encoder Representations from Transformers) in 2018 (Devlin et al., 2019). Since then, transformer-based models have been a widely used approach for several NLP tasks. Following, BERT has been outperformed by a variety of transformer-based architectures, including RoBERTa and DeBERTa, in several NLP tasks (Liu et al., 2019; He et al., 2021). In the specific domain of abusive speech, HateBERT has been introduced, showcasing better performance on abusive content than the original BERT model (Caselli et al., 2021).

BERT

While models before BERT still relied heavily on surface-level features, BERT introduced a richer, context-aware understanding of language (Devlin et al., 2019). This shift is especially important for complex NLP tasks like abusive speech detection, where nuance and context are critical.

BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. Unlike earlier approaches, which processed text either left-to-right or used shallow combinations of both directions, BERT fully integrates both contexts at every layer. As a result, it can be fine-tuned with just one additional output layer to achieve high performance across a wide range of tasks. Based on the Transformer architecture, BERT-base consists of 12 layers with a hidden size of 768, 12 self-attention heads, and 110 million parameters. Its pre-training involves a masked language model (MLM) objective, where some input tokens are randomly masked and predicted by the model. Additionally, BERT incorporates a next sentence prediction (NSP) task that jointly pre-trains text-pair representations (Devlin et al., 2019).

These characteristics make BERT easily adaptable to various NLP tasks with only minimal modifications. Due to its superior contextual understanding, the model has outperformed traditional models, such as bidirectional LSTMs, in detecting abusive speech, making it a strong candidate for abusive speech detection (Saleh et al., 2023).

HateBERT

One challenge with these pre-trained models is that the training language variety makes them well suited for general-purpose language understanding tasks, but they show limits with more domain-specific language varieties. This gave rise to several domain-specific BERT-like pre-trained language models, including HateBERT, which is a re-trained for abusive language detection.

The model was trained on RAL-E, a large-scale dataset of Reddit comments in English from communities banned for being offensive, abusive, or hateful (Caselli et al., 2021). With HateBERT, the authors kept the original model architecture and tokenizer unchanged. They pre-trained HateBERT using the MLM objective only, keeping the Next Sentence Prediction (NSP) task excluded. The final architecture of HateBERT is therefore identical to the BERT-base-uncased version (Caselli et al., 2021).

The new model showed better performance than the original BERT model when fine-tuned on several abusive speech datasets, including OffensEval, AbuseEval and HatEval (Caselli et al., 2021; Zampieri et al., 2019; Caselli et al., 2020; Basile et al., 2019). After that, the HateBERT model has been applied in various research, under which ToxiGen where it was used as a classifier and validation benchmark for their new dataset with a focus on subtle forms (Hartvigsen et al., 2022). This showed that HateBERT's domain-specific knowledge could help with the detection of abusive language, and in particular, implicit forms.

2.2.3 LLMs

More recently, the state-of-the-art has shifted toward larger, general-purpose language models, such as GPT, LLaMA or Qwen (OpenAI, 2024; et al., 2024; Yang et al., 2025). These models are trained on a vast amount of multilingual corpora and can be easily adapted to a variety of tasks via prompt engineering or more traditional fine-tuning.

Qwen

Among those models is Qwen, a family of large language models by Alibaba Cloud (Yang et al., 2025). The latest version of the model is 3.0, which was introduced in April 2025. As this was during the conduction of the current research, I will go in further depth on version 2.5, released in 2024 (Yang et al., 2025).

Qwen is a comprehensive series of LLMs that can be used for various tasks. Qwen 2.5 is pre-trained on 18 trillion tokens, and supervised and fine-tuned over 1 million samples, and further optimized through multistage reinforcement learning. The Qwen 2.5 LLM series comes in a rich set of configurations. The open-weight offerings include base models and instruction-tuned models in different parameter sizes varying from 0.5–72B, and quantized versions of the instruction-tuned models (Yang et al., 2025).

Even though the Qwen model is not originally built for abusive speech detection, it can be effectively applied to classification tasks in this area. Its open accessibility and support for instruction tuning make Qwen a practical model for research. Recent work has explored how LLMs can be applied to abusive speech detection and has shown promising results in prompt engineering for this task (Roy et al., 2023; Huang et al., 2023; Ziems et al., 2024).

Prompting LLMs

Prompt engineering is the practice of carefully designing prompts to guide model behavior effectively (Guo et al., 2023). By using this method, the effectiveness of an LLM can be significantly improved with little or no training, offering advantages over traditional fine-tuning.

The way a prompt is defined can vary in multiple ways. One such variation is the number of shots the model gets to see, referring to the number of examples provided. Han and Tang (2022) explored in their research the effect of the design of prompts for hate speech detection. Their results already showed very strong performance with zero-shot learning capabilities. Providing a few training examples (e.g., 8) in the prompt did not noticeably improve performance. Moreover, their results showed that only after the number of training examples reaches a decent amount (e.g., 16), model performance starts to improve. However, this boost does not continue to increase as the number of examples grows (Han and Tang, 2022). Therefore, they chose 16 as the ‘sweet spot‘ for the number of examples for their experiments.

The same authors also explored three different prompting techniques. First, instead of providing binary labels (e.g., hate speech or not), they further divided the labels into finer-level categories such as gender-offensive, race-offensive, etc. Second, instead of instructing the model to generate classification labels (either binary or multi-class), they tell the model to generate a Chain-of-Thought (CoT) before reaching a conclusion.

This obliges the model to create an extra thinking step. Third, they directly instructed the model about the perspectives to use in the task description, such as considering gender-offensive, race-offensive, etc. Their results showed that using more informative instructions, followed by adding CoT, seems to be the most effective way to inject prior knowledge into the model (Han and Tang, 2022).

Another study by Roy et al. (2023) explored different prompt variation techniques for hate speech detection on three large language models (including GPT-3.5) and three datasets: HateXplain, IHC, and ToxicSpans (OpenAI, 2023; Mathew et al., 2021; ElSherief et al., 2021; Pavlopoulos et al., 2021). The authors found that including target information improves model performance substantially (∼20–30%) over the baseline across the datasets. There is also a considerable effect from adding rationales or explanations (∼10–20%) over the baseline (Roy et al., 2023). These results show that including target information and rationales or explanations is an effective way to improve the prompt in hate speech detection.

The recent studies show that carefully crafting prompts for hate speech and abusive speech in general can effectively and easily improve the performance of LLMs, and that adding prior knowledge in a careful way through task phrasing, shots, and instructions is key to achieving this. Key strategies include using 16 shots, more informative instructions, and adding target information or explanations.

2.3 Summary

Related work has shown that determining what constitutes abusive speech is challenging due to overlapping definitions and inconsistencies in annotation guidelines. Abusive speech is commonly used as an umbrella term that includes both hate speech and offensive language. This definition will be adopted throughout the remainder of this research.

One of the challenges in detecting abusive speech involves the nuanced ways in which it can be expressed. Improving this is important for the overall detection of abusive speech. Over the last few years, several researchers have focused more on the distinction between explicit and implicit abuse to improve the overall detection of abusive speech (Caselli et al., 2020; ElSherief et al., 2021; Vidgen et al., 2021; Mathew et al., 2021; Hartvigsen et al., 2022). Despite previous efforts, the task remains challenging.

Since the early 2010s, abusive speech detection systems have evolved from rule-based approaches to large language models (LLMs) that are increasingly capable of detecting the nuance and complexity of language. In the last 5 years, the field has shifted with the introduction of BERT and domain-specific models such as HateBERT (Devlin et al., 2019; Caselli et al., 2021). Their contextual understanding has outperformed traditional models such as LSTMs.

More recently, the state-of-the-art has shifted toward larger, general-purpose models such as GPT, LLaMA, or Qwen (OpenAI, 2024; et al., 2024; Yang et al., 2025). These models are trained on vast multilingual corpora and can be easily adapted to a variety of tasks via prompt engineering or more traditional fine-tuning. The open accessibility and support for instruction tuning make models like Qwen a practical option for research (OpenAI, 2024).

Recent work has explored how LLMs can be applied to abusive speech detection and has shown promising results in prompt engineering for this task (Roy et al., 2023; Huang et al., 2023; Ziems et al., 2024). The way a prompt is defined can vary in multiple ways. Recent studies show that carefully crafting prompts for hate speech and abusive speech in general can effectively and easily improve the performance of LLMs, and that adding prior knowledge in a careful way through task phrasing, shots, and instructions is key to achieving this. Prior research has proven, in particular, that using 16 shots, more informative instructions, and adding target information or explanations are effective ways of prompting (Han and Tang, 2022; Roy et al., 2023).

Chapter 3

Dataset

The experiments in this study are primarily based on the AbuseEval v1.0 dataset, supplemented by the Implicit Hate Corpus (IHC) (Caselli et al., 2020; ElSherief et al., 2021). This chapter provides an overview of the details of these datasets. Both datasets are publicly available and widely used in research on abusive speech detection. They are particularly known for their focus on annotating implicit forms of abusive speech, making them a suitable match for this research.

To fine-tune the BERT models, I used both the AbuseEval and IHC training sets. I also used the AbuseEval training set to design the prompts for Qwen2.5. Evaluation of both the BERT models and the prompted Qwen2.5 was carried out on the AbuseEval test set. Table 3.2 summarizes the label counts and distribution across these datasets. Section 3.1 provides further details on AbuseEval, while Section 3.2 presents more information on the IHC dataset

Dataset	**Label**	**Training (No., %)**	**Test (No., %)**
AbuseEval	Explicit abuse	2,023 (15%)	106 (12%)
	Implicit abuse	726 (5%)	72 (8%)
	Not abuse	10,491 (79%)	682 (79%)
	Total	**13,240**	**860**
IHC	Explicit hate	1,089 (5%)	–
	Implicit hate	7,100 (33%)	–
	Not hate	13,291 (62%)	–
	Total	**21,480**	–
Total		**34,720**	**860**

Table 3.1: Distribution of labels in the AbuseEval and IHC datasets

3.1 AbuseEval

The AbuseEval dataset was introduced as the result of an effort to enrich and reannotate the OLID/OffensEval dataset originally introduced by ElSherief et al. (2021). The goal of creating AbuseEval was to move beyond surface-level cues and support the development of models that detect more subtle, context-dependent forms of abuse (Caselli et al., 2020). This makes this dataset particularly suitable for tasks focused on detecting implicit cases of abusive speech.

Dataset	Label	Training (No., %)	Test (No., %)
AbuseEval	Explicit abuse	2,023 (15.3%)	106 (12.3%)
	Implicit abuse	726 (5.5%)	72 (8.4%)
	Not abuse	10,491 (79.2%)	682 (79.3%)
	Total	**13,240**	**860**
IHC	Explicit hate	1,089 (5.1%)	–
	Implicit hate	7,100 (33.1%)	–
	Not hate	13,291 (61.9%)	–
	Total	**21,480**	–
Total		**34,720**	**860**

Table 3.2: Distribution of labels in the AbuseEval and IHC datasets

3.1.1 OLID/OffensEval

The original OLID/OffensEval dataset consists of 14,100 English tweets that were collected from Twitter in 2018 through keyword-based search, including a substantial number related to American politics. The dataset was initially annotated to (A) distinguish between *offensive* and *non-offensive* content. All instances labeled as *offensive* were further annotated to (B) identify targetness (*targeted* or *untargeted*). If an instance was classified as targeted, they additionally (C) specified the type of target (*individual*, *group*, or *other*) (Zampieri et al., 2019). Table 3.3 shows the counts and distribution of these labels. The dataset contains 13,240 training instances and 860 test instances, each annotated for subtask A, with subtasks B and C applied where relevant.

The original annotation labels for OLID/OffensEval are solely used for prompting and error analysis in the current study, and not for classification.

Subtask	Label	Train (No., %)	Test (No., %)
A	OFF (Offensive)	4,400 (33%)	240 (28%)
	NOT (Not Offensive)	8,840 (67%)	620 (72%)
	Total	**13,240**	**860**
B	TIN (Targeted)	3,876 (88%)	213 (89%)
	UNT (Untargeted)	524 (12%)	27 (11%)
	Total	**4,400**	**240**
C	IND (Individual)	2,407 (62%)	100 (56%)
	GRP (Group)	1,074 (28%)	78 (44%)
	OTH (Other)	395 (10%)	35 (20%)
	Total	**3,876**	**213**

Table 3.3: Distribution of labels in the original OLID/OffensEval dataset

3.1.2 Reannotation to AbuseEval

The newly created version, AbuseEval, was introduced by Caselli et al. (2020) to address some of the limitations of the original annotation, with a focus on the explicitness of the message and the role of context. The authors also chose the term abusive language as a better fit for the reannotated data than offensive language, as it covers a broader concept, aligning with discussions on terminology in abusive speech (see Subsection 2.1.1).

The AbuseEval dataset includes three categories: *explicit abuse*, *implicit abuse*, and *not abusive*. Examples of each class are provided in Table 3.4 below. This reannotation introduced a shift in label distribution. Table 3.2 shows a full overview of the new label counts. In the training set, 4,400 tweets (33%) were initially labeled as *offensive*. With the new annotation, 2,023 are labeled as *explicit abuse* and 726 as *implicit abuse*, meaning 2,749 tweets (20%) are now considered abusive. In the test set, 240 tweets (28%) were originally labeled as *offensive*. After reannotation, 106 are labeled as *explicit abuse* and 72 as *implicit abuse*, bringing the number of abusive tweets to 178 (20%). This shows an annotation shift from offensive to abusive for 1,651 tweets in the training set and 62 tweets in the test set.

In the current study, I used the AbuseEval training set for two purposes: training the BERT models and designing prompts for Qwen2.5. The test set was used for evaluating both approaches.

Label	**Tweet**
EXP	**Example 1**: @USER Liberals are all Kookoo !!! **Example 2**: @USER Canada doesn't need another CUCK! We already have enough #LooneyLeft #Liberals f**king up our great country! #Qproofs #TrudeauMustGo
IMP	**Example 1**: @USER was literally just talking about this lol all mass shootings like that have been set ups. it's propaganda used to divide us on major issues like gun control and terrorism **Example 2**: @USER @USER Kind of like when conservatives wanna associate everyone to their left as communist antifa members?
NOTABU	**Example 1**: @USER She should ask a few native Americans what their take on this is. **Example 2**: @USER @USER Go home you're drunk!!! @USER #MAGA #Trump2020 URL

Table 3.4: Example tweets per label of AbuseEval v1.0 (EXP = Explicit Abuse, IMP = Implicit Abuse, NOTABU = Not Abusive)

3.2 Implicit Hate Corpus (IHC)

Similar to AbuseEval, the Implicit Hate Corpus (IHC) by ElSherief et al. (2021) was created to better understand and detect implicit forms of *hate speech*. With this, the dataset provides a useful foundation for studying hate speech that is less explicit. In my study, I used only the training set of IHC as additional data to train the BERT models. Although IHC focuses specifically on hate speech and not all forms of abusive language, hate speech is a subset of abusive speech and thus complements the existing data.

The training set of IHC consists of 21,480 tweets associated with prevalent hate groups in the United States, collected between 2015 and 2017 (ElSherief et al., 2021). The authors distinguish between *explicit hate*, *implicit hate*, and *not hate* in their annotation. Table 3.2 provides a full overview of the label count and distribution. Out of all tweets, 7,100 (33%) were identified as containing implicit hate speech, which is substantially higher than the proportion reported in earlier work. Table 3.5 illustrates an overview of examples.

Label	Tweet
explicit_hate	**Example 1**: rt crooked hillary you didn't lose because you couldn't portray yourself as an " agent of change " " you loss because you are a socialist agent of destruction . shut up hillary and stop whining ! " **Example 2**: rt estonian pm : " i don't want blacks in estonia ; this is a white country " " via ¡ well said . "
implicit_hate	**Example 1**: " : jewish harvard professor noel ignatiev wants to abolish the white race via #wr " " " **Example 2**: rt : tune in today's jmt for my interview w / robert spencer on " the complete infidel's guide to iran ! " " "
not_hate	**Example 1**: b.higher education is a part of european culture that was imported to other continents although the asian culture pressures their people to aim to be the best at everything (hence all the stereotypical jokes) in asian culture " studying is everything " " " **Example 2**: has a problem with " the whites " " and " " the christians . " " go ahead you are free to say it now ! "

Table 3.5: Example tweets per label of IHC (explicit_hate = Explicit Hate, implicit_hate = Implicit Hate, not_hate = Not Hate)

The authors of IHC also introduced a six-class taxonomy for characterizing and detecting different forms of implicit hate, grounded in social science and relevant NLP literature (ElSherief et al., 2021). In the current study, these labels are used solely for prompt design and not for classification purposes. ElSherief et al. (2021) state the following six categories:

- **White Grievance:** Frustration over a minority group's perceived privilege and casting majority groups as the real victims of racism. Linked to extremist behavior and support for violence.
- **Incitement to Violence:** Flaunting ingroup unity and power or elevating known hate groups and ideologies. Speech inciting violence is prohibited by law.
- **Inferiority Language:** Implies one group or individual is inferior, including dehumanization and toxification, both early warning signs of genocide. Related to assaults on human dignity, dominance, and declarations of in-group superiority.
- **Irony:** Use of sarcasm, humor, and satire to attack or demean a protected class or individual. Commonly used by hate groups to mask hatred and extremism.
- **Stereotypes and Misinformation:** Associates a protected class with negative attributes such as crime or terrorism. Includes misinformation that feeds stereotypes and vice versa, like Holocaust denial.
- **Threatening and Intimidation:** Conveys a commitment to a target's pain, injury, or violation of rights, including subtle forms of intimidation.

Chapter 4

Methodology

This chapter outlines the methodological approach behind the experiments performed in this research. To explore the detection of abusive speech, I conducted two main experiments — fine-tuning BERT-based models and prompting a generative model — and applied both to binary and ternary classification tasks. Section 4.1 describes the fine-tuning process applied to BERT and HateBERT (Devlin et al., 2019; Caselli et al., 2021). Section 4.2 covers the prompt engineering experiments conducted on Qwen2.5. Finally, Section 4.3 provides an overview of the metrics used for evaluation of the models.

4.1 Fine-tuning (Hate)BERT

Prior work has shown that the contextual understanding of the general-purpose BERT, as well as its domain-specific variant trained on abusive speech, HateBERT, has been effective in detecting abusive content (as discussed in subsection 2.2.2). Moreover, the pre-training with domain-specific knowledge done with HateBERT has been shown to be beneficial for identifying more nuanced and implicit forms of abuse.

Therefore, in the first set of experiments, I evaluate the performance of both BERT and HateBERT in abusive speech detection. This is done by conducting four experimental setups, each applied using both the BERT-base-uncased model and HateBERT. An overview of these setups is provided in Section 4.1.

Model	**Training Data**	**Binary**	**Ternary**
BERT	AbuseEval	x	x
	AbuseEval + IHC	x	x
HateBERT	AbuseEval	x	x
	AbuseEval + IHC	x	x

Table 4.1: Overview of experimental setups for BERT and HateBERT.

I used two different training datasets: the AbuseEval training set alone, and a combination of the AbuseEval and IHC dataset to test the effect of additional implicit abusive speech data on the performance of the models. I fine-tuned the model on two different classification schemes: binary classification (*abusive* vs. *not abusive*) and ternary classification (*explicit abuse*, *implicit abuse*, *not abusive*). For the binary classification setup, I merged the original *implicit abuse* and *explicit abuse* labels into a single *abusive* label. This resulted in a total of four experiments (two models × two datasets) per classification scheme, where I evaluated each model on the AbuseEval test set.

Together, these experiments provide a solid foundation for evaluating the impact of using a domain-specific model, as opposed to a general-purpose one, and including additional data in the detection of abusive speech, particularly in implicit form. The best model will be configured as a comparison for the prompting experiments.

4.1.1 Training

To ensure a consistent experimental setup, I carried out the same process for all four models. This enabled a controlled evaluation of the effects of both model type and training data. Crucial in the training process was addressing the class imbalances of the dataset.

During data preparation, I divided the training set into training and validation sets with a 90/10 split. The test set was kept separate and was not used during training or validation. I applied stratified sampling, which ensures that the class distribution is maintained in both train and validation splits. The data was fed to the model by tokenizing the tweets for sequence classification with a maximum sequence length of 128 tokens, which covered 99.87% of the data without requiring truncation. The original tokenizer of each model was used, with padding and truncation applied to the maximum sequence length.

To further address class imbalance, I used a weighted cross-entropy loss, which assigns a different weight to each class. I calculated balanced class weights based on the training labels and adapted HuggingFace's `Trainer`[1] to incorporate these weights. This allowed the model to give proportionally more attention to underrepresented classes. I also used a fixed random seed (42) to ensure reproducibility in both training and evaluation. To mitigate overfitting, I applied L2 regularization with a weight decay of 0.01. The other training hyperparameters included a learning rate of 2e-5, a batch size of 32, evaluation after each epoch, and early stopping with a patience of 2.

Based on validation performance, I saved the best-performing checkpoints, resulting in four models for each classification scheme that were then evaluated on the held-out AbuseEval test set.

4.2 Prompting Qwen2.5

Previous work has exposed that applying prompt engineering to LLMs can be an effective approach for detecting abusive speech, and that carefully crafting a prompt through task phrasing, shots, and instructions is key to achieving this (as discussed in subsection 2.2.3). Therefore, the second set of experiments involved prompting the quantized version of Qwen2.5, qwen2.5-7b-instruct (Yang et al., 2025). The 7b-instruct version is large enough to show strong language understanding and small enough to run locally. These experiments also covered both binary classification (*abusive* vs. *not abusive*) and ternary classification (*explicit abuse*, *implicit abuse*, *not abusive*).

Table 4.2 provides a summary of the configuration steps for these experiments. The experiments involve a 3-step set-up: **(I)** For each strategy (binary and ternary), I tested four variations of shots, including zero-shot and few-shot settings with different

[1] `https://huggingface.co/docs/transformers/main_classes/trainer`

[1] Temperature values other than 0.0 were tested only for the best-performing setups.

numbers of examples per class. **(II)** I explored five different prompting strategies to examine their impact on the classification performance of the model. This resulted in a total of twenty experiments (four shot quantities x five strategies) per classification scheme. All experiments were initially conducted with a default temperature of 0 and evaluated on the AbuseEval test set. **(III)** In the final step, I re-evaluated the setups that performed best in step II by testing three additional temperature values.

I. Shots/Class	**II. Prompting Strategies**	**III. Temperature**[2]
Binary: 0, 1, 8, 12	Base	0.0
Ternary: 0, 1, 6, 8	Definition	0.3
	Chain-of-Thought (CoT)	0.5
	CoT with Targeting	0.8
	CoT with IHC Labels	

Table 4.2: Overview of setups for prompting experiments

This approach enables a controlled experimental setup to investigate the effects of prompt strategy, shot quantity, and temperature on the classification. The following subsections will explain each step and its underlying motivation in further detail.

4.2.1 Shots

Experiments in related works, as discussed in subsection 2.2.3, have shown that example-label pairs achieved very strong performance in zero-shot settings, and the model's performance only started to improve after the number of training examples reached a decent amount (e.g., 16). After that, the boost did not continue as the number of examples increased (Han and Tang, 2022).

Building on these conclusions, I selected four different configurations per classification scheme for testing. These setups are outlined in Table 4.3 below. For the binary classification setup, I used 0, 1, 8, and 12 shots per class, resulting in a total of 0, 2, 16, and 24 examples. For the ternary classification setup, I used 0, 1, 6, and 8 shots per class, resulting in a total of 0, 3, 18, and 24 examples. Based on prior findings, we would expect performance to be strongest at 0 shots (due to strong zero-shot capabilities) and at 16/18 total shots (the suggested 'sweet spot' in performance). The 2/3 and 24 total shot configurations are included as additional experimental reference points.

Setup	**Shots per class**	**Total shots**
Binary	0	0
	1	2
	8	16
	12	24
Ternary	0	0
	1	3
	6	18
	8	24

Table 4.3: Shots per class and total number of examples.

I used a systematic approach to choose the examples for the shots, to ensure they were both balanced and effective. I selected the examples from the tweets in the AbuseEval training set, based on three criteria: **(A)** targetness, using the OLID/OffensEval subtask labels (Zampieri et al., 2019); **(B)** balanced topic coverage, to reduce the dataset's bias toward political content; and **(C)** linguistic features, based on related work and an evaluation of the training set. Based on these criteria, I selected a representative example set.

(A) Targetness

In the first check, I used the original OLID/OffensEval subtask labels for targetness to ensure a structured and balanced distribution of examples that both included and did not include a target (Zampieri et al., 2019). The dataset description in section 3.1 provides more detail about the subtask labels. This resulted in an example set where each label was equally represented. These labels were:

- Targeted to an Individual **(TIN/IND)**
- Targeted to a Group **(TIN/GRP)**
- Targeted to Other **(TIN/OTH)**
- Untargeted **(UNT)**

(B) Topic

Secondly, I manually reviewed the examples to ensure topic diversity. Since AbuseEval contains a relatively high proportion of tweets focused on American politics, including political insults, I made sure to select a broader range of themes to reduce topical bias. These included racism, mental health–related abuse, body shaming, gender shaming, and slut-shaming.

(C) Linguistic Features

Lastly, I compiled a set of linguistic features for each category, *explicit abuse*, *implicit abuse*, and *not abusive*, based on previous research on these categories (as discussed in subsection 2.1.2) and manually identified patterns in the training data. These features served as a checklist to help ensure the selected examples captured the most common characteristics of each category. I then checked if all features are represented in the examples.

Explicit Abuse	**Implicit Abuse**	**Not Abusive**
Profanity	Limited or no profanity	Politeness
Direct pronouns and names	Stereotyping and generalizations	Empathy
Uppercase words and exclamations	Rhetorical questions	Neutral or positive tone
Syntactic simplicity	Complex or longer words	Syntactic complexity
Imperative structure	Coded or euphemistic language	Constructive critique
Direct insults, incl. slurs	Sarcasm and irony	
	False politeness / passive-aggressiveness	

Table 4.4: Linguistic features across abuse types.

Shots selection

This resulted in the full balanced order of examples showcased in Table 4.5. This table shows the examples for the ternary setup. The examples are labeled with their label (EXP, IMP, or NOTABU), fine-grained subclass A (TIN or UNT), subclass B (IND, GRP, or OTH), and topic. In the binary setup, I combined the explicit (EXP) and implicit (IMP) class into the abusive (ABU) class. The full prompts, including the binary distribution, can be found in Appendix 8.

Depending on the number of shots used per setup, I sampled the examples in a stratified manner to preserve the balance in targetness, and additionally, topic and linguistic features as much as possible.

To illustrate how this works, I take the ternary 18-shot setup (6 per class) as an example. In this case, the system selects the first 6 examples from the EXP list, the first 6 from the IMP list, and the first 6 from the NOTABU list. The selected examples from the EXP and IMP lists include: TIN / IND, TIN / GRP, etc., and for IMP: UNT, TIN / OTH, etc. In other words: 3 x TIN / IND, TIN / GRP, TIN / OTH, and UNT, balanced over 6 EXP and 6 IMP examples. This is applicable to every number of shots: the subclass label distribution remains balanced. This ensures a consistent set-up across different numbers of shots.

4.2.2 Prompting Strategies

There are various ways of prompting. Prior research has shown that adding prior knowledge in a careful way through task phrasing, shots, and instructions is key to achieving higher performance with LLMs (as discussed in section 4.2). Previous work showed that adding informative instruction and requesting a CoT are among the most effective prompting techniques. Other research also showed that including target information or explanations is beneficial for the model's performance.

To evaluate how different types of prompting affect model performance, I designed and tested five prompting strategies based on this prior work, outlined below. These strategies range from minimal instructions to more guided prompts. Strategy **1** serves as the baseline. Strategy **2** tests the effect of adding a definition to the prompt. Strategies **3–5** are based on prior work (CoT, target information, and informative instructions).

1. **Base:** Classification request without any definitions or guidance.
   Example: "Classify the following text into one of the following categories: - explicit abuse (EXP); - implicit abuse (IMP); not abusive (NOTABU)"

2. **Definition:** Classification request with definitions of the categories provided.
   Example: "Classify the following text into one of the following categories: - explicit abuse (EXP): Language with direct and literal forms of abusive speech, such as slurs, profanity, and other clearly hostile or offensive expressions."

3. **Chain-of-Thought (CoT):** Classifcation request with a short explanation of reasoning.
   Example: "Write 1–2 short sentences explaining why the text is classified as EXP or IMP or NOTABU."

4. **Chain-of-Thought with Targeting:** The same as **CoT**, but explicitly asks the model to consider whether the text is targeted or untargeted.
   Example: "Consider in your explanation if the text is targeted or untargeted."

5. **Chain-of-Thought with IHC Labels:** The same as **CoT**, but explicitly asks the model to consider six implicitness categories based on the taxonomy by ElSherief et al. (2021).
   Example: "Consider in your explanation whether the text includes any of the following: - white grievance; - incitement to violence; - inferiority language; - irony; - stereotypes and misinformation; - threatening and intimidation"

4.2.3 Temperature

During the first round of experiments, I used a default temperature setting of 0.0 in order to minimize output variability and isolate the effect of prompt structure. After this initial evaluation, I re-evaluated the best-performing prompts on the AbuseEval test set using higher temperature settings: 0.3, 0.5, and 0.8.

4.3 Evaluation

For the evaluation of all experiments, I used the `classification_report` function from *scikit-learn*, which reports three key metrics: precision, recall, and F-score (Pedregosa et al., 2024).

- **Precision (P)** measures the accuracy of the model's positive predictions. It is calculated as the ratio of true positives (TP) to the total number of predicted positives:

$$Precision = \frac{TP}{TP + FP}$$

- **Recall (R)** measures the model's ability to capture all actual positive instances. It is defined as the ratio of true positives to the total number of actual positives:

$$Recall = \frac{TP}{TP + FN}$$

- **F1-score (F1)** is a single measure that trades off precision versus recall into the weighted harmonic mean (Manning et al., 2008). There is often a trade-off between precision and recall, where improving one leads to a decrease in the other. The F1-score gives an overall picture of the model's performance:

$$F - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

The reported averages include **macro average** (averaging the unweighted mean per label), **weighted average** (averaging the support-weighted mean per label), and **sample average** (only for multilabel classification). To compare the evaluation across different models, I employed macro-averaging. Macro-averaging calculates metrics for each label and finds their unweighted mean, treating all classes equally, regardless of their frequency (Pedregosa et al., 2024).

#	Label	A	B	Topic	Tweet
Explicit Abuse (EXP)					
1	EXP	TIN	IND	Politics	@USER you are a lying corrupt traitor!!! Nobody wants to hear anymore of your lies!!! #DeepStateCorruption URL
2	EXP	TIN	GRP	Fat shaming	@USER 'the sex symbol of rap' LMAOOO BITCH YOUre FAT aS FUCK ANYWAYS you tryna fight?
3	EXP	TIN	OTH	Politics	MAGA are PERVERTS AND DEMONS—PERIOD! NOTHING IS SACRED OR INNOCENT WITH THESE MUTTS OF SATAN URL
4	EXP	UNT	–	Activity	@USER What the fuck game are you watching?
5	EXP	TIN	IND	Gender	@USER If you go by anything other than he or she you are fucked in the head.
6	EXP	TIN	GRP	Politics	@USER @USER The FASCISTS are Anti-American Communist Scum like yourself!! Now go fuck yourself you piece of human garbage!! USA #MAGA ANTIFA are Terrorists
7	EXP	TIN	OTH	Mental health	@USER @USER Are you bipolar? You keep saying different shit to people commenting on what u said
8	EXP	UNT	–	Insult	@USER So what ur saying is that it is in fact chicken shit and you are eating chicken shit URL
Implicit Abuse (IMP)					
1	IMP	UNT	–	Intelligence	@USER Gun Control means using both hands. And in this case means acquiring the correct target. You can't fix stupid.
2	IMP	TIN	OTH	Racism	10kids from different women from this American jobless men? Nigerian men you are the bomb!!!
3	IMP	TIN	GRP	Politics	4 out of 10 British people are basically full-on racists. 4 out of 10 voters vote for the Conservatives. Coincidence!???!???
4	IMP	TIN	IND	Politics	@USER Thank you for your fortitude and perseverance. Please give McConnell a kick in the butt from some of us conservatives.
5	IMP	UNT	–	Gender	2 hours at midnight and she said she is not a call girl? -Awkward!
6	IMP	TIN	OTH	Racism	@USER Do you think she is afraid of pigs that are not the same colour as her?
7	IMP	TIN	GRP	Fat shaming	@USER Love how fat they all are while chowing down on cake and cookies. That's the future liberals want.
8	IMP	TIN	IND	Politics	@USER Absolutely should be a nail in the coffin.... for Hillary
Not Abusive (NOTABU)					
1	NOTABU	–	–	Politics	@USER What if knowing that abortion is at stake in November mobilizes pro life forces? Kavanaugh being denied makes it critical for conservatives to maintain the Senate. Because most (R) seats are not up for election it could be a blessing.
2	NOTABU	–	–	Human rights	(You are! My prayer is that you are blessed with someone you know like that also. Someone who feels your basic human rights are worth protecting. Especially your God given right to personal privacy.)
3	NOTABU	–	–	Education	@USER You should reach out to Paul.berger@USER He is teaching & researching these classes. & he is fierce & great! V. Ethical.
4	NOTABU	–	–	Censorship	@USER You guys have featured splendidrainco's etsy page and she is being censored by etsy now. You can't find her site through the search engine you have to have a direct link. And according to her instagram post they have been basically leaving her out to dry.
5	NOTABU	–	–	Compliment	@USER Aww Bradley you are so cute
6	NOTABU	–	–	Health	@USER So happy for you and your dad!!! My dad had the same kind of cancer but stage 4. He was diagnosed in 2016 and passed away this passed January. My mom is kicking leukemia's ass since Nov. of 2015 and she is doing awesome!
7	NOTABU	–	–	Sports	@USER @USER It's OK that you are confused. Paul Guenther is a proven DC and play caller. We are rebuilding and trying to remain competitive at the same time. We will have steady improvement
8	NOTABU	–	–	Entertainment	@USER I loved the House Bunny. And the remembering names voice was funny. My family and I watch Mom it's great. I'm a recovered addict or moderately recovered if you are one who counts pot lol. It's a very good show.

Table 4.5: Prompt examples for ternary set-up labeled by label, subtasks, and topic.

Chapter 5

Results

This chapter presents an overview of the results of the conducted experiments. Section 5.1 describes the results of the fine-tuning experiments on the BERT-based models. Section 5.2 covers the outcomes of the prompting experiments conducted on Qwen2.5. Each section includes both the binary and ternary classification evaluation.

5.1 Fine-tuning (Hate)BERT

The first set of experiments includes fine-tuning BERT-based models. For each classification set-up, binary and ternary, I conducted four experiments. These involved fine-tuning both BERT and HateBERT on the AbuseEval dataset alone, as well as on the combined AbuseEval + IHC dataset.

I will first discuss the evaluation for each classification scheme and then provide an overall summary.

5.1.1 Binary classification

In the binary classification setup, the model's task is to distinguish between abusive (ABU) and non-abusive (NOTABU) speech. Table 5.1 provides an overview of the results for all four setups in binary classification. The rows list the two models (BERT and HateBERT), while the columns under the "Training Data" header show the performance on both the AbuseEval dataset alone and the combined AbuseEval + IHC dataset.

Model	**Label**	**Training Data**	
		AbuseEval	**AbuseEval + IHC**
BERT	ABU	0.63	0.57
	NOTABU	0.90	0.88
	Macro F1	0.76	0.72
HateBERT	ABU	0.64	0.55
	NOTABU	0.91	0.87
	Macro F1	**0.77**	0.71

Table 5.1: Binary classification performance of BERT and HateBERT.

Examining the models, HateBERT slightly outperforms the original BERT model with an overall performance of 0.77 for HateBERT vs. 0.76 for BERT on the AbuseEval dataset alone. On the combined set, the other way around occurs: BERT slightly outperforms HateBERT with 0.72 for BERT vs. 0.71 for HateBERT. This indicates that the models perform similarly on the datasets and the effect of domain-specific pretraining is limited.

When analyzing the effect of the dataset, we see that both BERT and HateBERT models trained solely on AbuseEval outperform those trained with the added IHC dataset (0.76 vs. 0.72 for BERT; 0.77 vs. 0.71 for HateBERT). These results suggest that adding more varied data does not necessarily improve generalization in binary classification. Notably, this better performance could be caused by the data being more similar to what the model was trained on. In the solo setup, both the training and testing sets come from the same dataset, AbuseEval. So the performance may be influenced by overfitting to the training set.

The model with the best overall performance for binary classification is HateBERT, trained solely on the AbuseEval training set. However, it only slightly outperforms BERT in this case, indicating that the impact of domain-specific pretraining is limited. The results also demonstrate better performance on the solo, less varied dataset, though it is important to note that the training and test sets share similar characteristics, which may have influenced the outcome.

5.1.2 Ternary classification

In the ternary classification setup, the task of the model is to distinguish between explicit abuse (EXP), implicit abuse (IMP), and non-abusive (NOTABU) speech. Table 5.2 showcases an overview of the results for all four setups in ternary classification. Similar to the binary classification setup, the rows list the two models, while the columns under ”Training Data” show performance on both the AbuseEval dataset alone and the combined AbuseEval + IHC dataset.

Model	**Label**	**Training Data**	
		AbuseEval	**AbuseEval + IHC**
BERT	EXP	0.65	0.63
	IMP	0.20	**0.29**
	NOTABU	0.90	0.88
	Macro F1	0.58	**0.60**
HateBERT	EXP	0.63	0.56
	IMP	0.20	0.26
	NOTABU	0.90	0.88
	Macro F1	0.57	0.57

Table 5.2: Ternary classification performance of BERT and HateBERT.

Analyzing the models, BERT slightly outperforms HateBERT on both the solo training data (0.58 for BERT vs. 0.57 for HateBERT) and the combined dataset (0.60 for BERT vs. 0.57 for HateBERT). This suggests that, for ternary classification, using a general model is more effective than relying on domain-specific pretraining.

The effect of the dataset choice is reflected in the slight improvement in overall performance for BERT when trained on the combined dataset (0.58 vs. 0.60). For HateBERT, overall performance remains unchanged (0.57). The most notable improvement for both BERT and HateBERT appears in identifying implicit abuse when trained on the combined dataset instead of the solo set (0.20 vs. 0.29 for BERT; 0.20 vs. 0.26 for HateBERT). Nevertheless, the scores for implicit abuse remain remarkably lower than those for the other two labels, highlighting the challenges the model encountered when detecting implicit forms.

The model with the best overall performance for ternary classification is BERT trained on the combined training set, emphasizing the relevance of using a general model together with more diverse data. Although the scores for implicit abuse remain low, this category benefits the most from the BERT model in combination with the combined dataset.

5.1.3 Summary

The results show that the binary and ternary classification tasks are differently affected by model and training data choices.

For binary classification, HateBERT trained solely on the AbuseEval dataset achieves the highest performance, suggesting that (although minimal) domain-specific pretraining can be beneficial. The model benefits the most from the solo dataset, which is likely affected by domain consistency in both the training and test sets.

In contrast, BERT trained on the combined dataset shows the best performance for ternary classification. While the scores for implicit cases remain lower than for other categories, this category shows the most notable improvement when using the combined dataset. This highlights both the challenges of detecting implicit abuse and the value of more varied training data in helping the model generalize to such cases.

5.2 Prompting Qwen2.5

The second set of experiments includes prompting Qwen2.5 for the two classification setups (binary and ternary).

The prompting experiments are divided into two rounds. The first round consists of 5 prompting experiments: (1) Baseline (**Base**), (2) Definition (**Def**), (3) Chain-of-Thoughts (**CoT**), (4) Chain-of-Thoughts with Targeting (**Tar**), and (5) Chain-of-Thoughts with IHC labels (**IHC**). All experiments were executed with a number of shots of 0, 1, 8, and 12 for binary and 0, 1, 6, and 8 for ternary, and a temperature of 0.0. Resulting in 20 experiments per classification scheme. The experiments with the best performance were then subjected to a second round that included a different range of temperatures: 0.2, 0.5, and 0.8. An in-depth explanation of the experiments is provided in Section 4.2.

I will first look at the results per classification setup and then discuss an overall evaluation.

5.2.1 Binary classification

We begin by examining the binary classification task, which distinguishes between abusive speech (ABU) and non-abusive speech (NOTABU). Table 5.3 provides an overview of the results for the first round of experiments that included the 20 experiments.

Shots	**Label**	**Experiment**				
		Base	**Def**	**CoT**	**Tar**	**IHC**
0	ABU	0.54	0.54	0.53	0.53	0.49
	NOTABU	0.83	0.79	0.84	0.86	0.86
	Macro F1	0.69	0.67	0.69	0.70	0.67
1	ABU	0.55	0.54	0.54	0.52	0.53
	NOTABU	0.85	0.82	0.86	0.85	0.86
	Macro F1	0.70	0.68	0.70	0.68	0.69
8	ABU	0.53	0.56	0.57	0.53	0.54
	NOTABU	0.79	0.82	0.83	0.84	0.81
	Macro F1	0.66	0.69	0.70	0.69	0.68
12	ABU	0.51	0.53	0.56	0.58	0.54
	NOTABU	0.77	0.80	0.86	0.84	0.81
	Macro F1	0.64	0.67	0.71	0.70	0.68

Table 5.3: Binary classification performance across different prompting strategies with temperature set to 0.0. **Base** = baseline; **Def** = definition; **CoT** = Chain-of-Thought; **Tar** = Chain-of-Thought with targeting; **IHC** = Chain-of-Thought with IHC labels. Green indicates improvements over Base (0-shot).

Experiment	**Shots**	**Label**	**Temperature**			
			0.0	**0.2**	**0.5**	**0.8**
CoT	1	ABU	0.54	0.54	0.55	0.57
		NOTABU	0.86	0.86	0.86	0.87
		Macro F1	0.70	0.70	0.71	0.72
	8	ABU	0.57	0.52	0.54	0.55
		NOTABU	0.83	0.82	0.82	0.82
		Macro F1	0.70	0.67	0.68	0.69
	12	ABU	0.56	0.56	0.57	0.54
		NOTABU	0.86	0.85	0.86	0.85
		Macro F1	0.71	0.71	0.72	0.69
Tar	0	ABU	0.53	0.55	0.52	0.55
		NOTABU	0.86	0.87	0.86	0.86
		Macro F1	0.70	0.71	0.69	0.71
	12	ABU	0.58	0.53	0.52	0.54
		NOTABU	0.84	0.85	0.85	0.86
		Macro F1	0.70	0.69	0.69	0.70

Table 5.4: Binary classification performance across different temperatures. Green indicates improvements over temperature of 0.0.

Examining the five experimental setups, we observe variation in performance across them. Prompting strategies that incorporate CoT or CoT with targetness tend to outperform the baseline score of 0.69. On the other hand, adding definitions or IHC labels seems to decrease performance in the binary classification task.

When analyzing shot configurations, I observed some performance improvements with higher shot counts, particularly in the CoT experiments. The CoT setup with 12 shots achieved the highest initial performance of 0.71.

The five best-performing experiments were then tested under varying temperature settings in the second round, as shown in Table 5.4. Performance improved slightly for both CoT (1 shot and 12 shots) and Tar (0 shot) configurations at higher temperatures. The best overall result was achieved by the CoT (1 shot) setup with a temperature of 0.8, reaching a score of 0.72.

5.2.2 Ternary classifcation

Next, I examined the ternary classification results, which includes the distinction between explicit abuse (EXP), implicit abuse (IMP), or non-abusive (NOTABU). Table 5.5 provides an overview of the results for the first round of experiments that included the 20 experiments.

Looking at the different experiments, we see improvement over the baseline across most of them. Similar to the binary results, CoT and CoT with targetness achieve the highest performances, while Def and IHC show the lowest overall results.

Analyzing the effect of shots, we see especially the impact of shots on the overall performance in baseline, with the scores going up from 0.47 (0-shot) to 0.50 (8-shot). In the other categories, the effect is more mixed. In the end, the Tar (0-shots) setup achieved the highest overall performance of 0.52, followed by CoT (8-shots).

While IHC did not improve overall performance, it did significantly increase the implicit score at the expense of the explicit class. Both Base (8-shots), Tar (0-shots), and IHC (0-shots) show notable improvement in the implicit class, with scores of 0.25, 0.25, and 0.26 over the baseline of 0.19. This suggests that either increasing the number of shots or adding guided reasoning can improve the detection of implicit cases significantly, but the combination does not seem effective.

The three best-performing experiments were then tested with different temperature settings in a second round. Table 5.6 shows the results of this. Only the performance of CoT with targetness improved at a temperature of 0.20, resulting in the best overall performance of 0.53.

5.2.3 Summary

Across both binary and ternary classification setups, we see that incorporating CoT or CoT with targetness is most effective in improving prompt performance, while Def and IHC have the least impact on overall scores. This shows that adding a (guided) reasoning step is beneficial for the detection of abusive speech.

In the binary task, CoT (12-shot) achieved the highest performance of 0.71 in the initial round with the default temperature of 0.0. In the second round, CoT (1-shot) with a temperature of 0.8 yielded the best overall score of 0.72. In the ternary setup, Tar (0-shot) achieved the highest performance with an F1 score of 0.52 in the first round, which increased slightly to 0.53 in the second round at a temperature of 0.2. This suggests only a minimal effect of temperature changes on the scores.

Shots	Label (F1)	Experiment				
		Base	**Def**	**CoT**	**Tar**	**IHC**
0	EXP	0.42	0.32	0.46	0.43	0.34
	IMP	0.19	0.18	0.16	0.25	0.26
	NOTABU	0.82	0.72	0.84	0.86	0.85
	Macro F1	0.47	0.40	0.49	0.52	0.48
1	EXP	0.43	0.42	0.43	0.40	0.41
	IMP	0.22	0.19	0.18	0.15	0.21
	NOTABU	0.83	0.79	0.84	0.84	0.84
	Macro F1	0.49	0.47	0.48	0.46	0.49
6	EXP	0.43	0.42	0.45	0.46	0.40
	IMP	0.23	0.22	0.14	0.22	0.22
	NOTABU	0.83	0.80	0.81	0.80	0.80
	Macro F1	0.50	0.48	0.47	0.49	0.47
8	EXP	0.42	0.43	0.48	0.45	0.41
	IMP	0.25	0.22	0.21	0.17	0.21
	NOTABU	0.82	0.79	0.82	0.80	0.79
	Macro F1	0.50	0.48	0.51	0.47	0.47

Table 5.5: Ternary classification performance across different prompting strategies with temperature set to 0.0. **Base** = baseline; **Def** = definition; **CoT** = Chain-of-Thought; **Tar** = Chain-of-Thought with targeting; **IHC** = Chain-of-Thought with IHC labels. Green indicates improvements over Base (0-shot).

Experiment	Shots	Label	Temperature			
			0.0	**0.2**	**0.5**	**0.8**
Base	8	EXP	0.42	0.42	0.43	0.45
		IMP	0.25	0.23	0.24	0.20
		NOTABU	0.82	0.82	0.82	0.82
		Macro F1	0.50	0.49	0.49	0.49
CoT	8	EXP	0.48	0.43	0.47	0.45
		IMP	0.21	0.20	0.17	0.17
		NOTABU	0.82	0.82	0.83	0.82
		Macro F1	0.51	0.48	0.49	0.48
Tar	0	EXP	0.43	0.47	0.44	0.42
		IMP	0.25	0.24	0.17	0.24
		NOTABU	0.86	0.87	0.85	0.86
		Macro F1	0.52	0.53	0.48	0.51
IHC	0	EXP	0.34	0.30	0.33	0.29
		IMP	0.26	0.21	0.20	0.19
		NOTABU	0.85	0.85	0.84	0.84
		Macro F1	0.48	0.45	0.46	0.44

Table 5.6: Ternary classification performance across different temperatures. Green indicates improvements over temperature of 0.0.

For the implicit class, Base (8-shots), Tar (0-shots), and IHC (0-shots) show similar scores of 0.25, 0.25, and 0.26. This shows that either higher shots or guided reasoning can help improve the detection of implicit cases, but not necessarily the combination of the two.

In conclusion, the results show that CoT and CoT with targetness perform best. This shows that the model performs particularly well when its own knowledge is activated through reasoning, and that adding a (guided) reasoning step is beneficial for the detection of abusive speech. It also suggests that, when tested on the AbuseEval dataset, the model already has a solid understanding of what implicit abuse, explicit abuse, and non-abusive language look like.

Chapter 6

Error Analysis

To better understand where the model performs well and where it misclassifies, I subjected the Qwen2.5 model, with the prompt strategy CoT Targeted with 0 shots setting in the ternary set-up, to an error analysis. This version of the model showed the overall best performance when comparing the models in a default setting of temperature 0.0. Adjusting the temperature to 0.2 had minimal impact on performance and did not provide any clear advantages over using a deterministic setting. Therefore, the version with a temperature of 0.0 is used for the error analysis.

Section 6.1 discusses the classification report and outlines the specific confusion between labels. After that, I dive deeper into the patterns in the errors, specifically related to implicit instances. Section 6.2 describes the manual part of the analysis, focusing on linguistic patterns, while Section 6.3 covers the automated part of the analysis.

6.1 Overall performance

6.1.1 Classification Report

To gain a clearer understanding of the model's performance across the three classes, I first examined the full classification report, shown in Table 6.1.

Label	**Precision**	**Recall**	**F1-score**	**Support**
EXP	0.46	0.41	0.43	106
IMP	0.20	0.35	0.25	72
NOTABU	0.90	0.84	0.86	682
Micro Avg	0.74	0.74	0.74	860
Macro Avg	0.52	0.53	0.52	860
Weighted Avg	0.78	0.74	0.76	860

Table 6.1: Classification report for Qwen2.5 using the CoT Targeted prompting strategy (0-shot).

The model's overall performance is moderate, with a macro-average F1-score of 0.52. The higher micro-average of 0.74 reflects the influence of the more frequent occurring class NOTABU. The model also performs best on this class, achieving an F1-score of 0.86. Both precision (0.90) and recall (0.84) are strong, indicating that the model correctly identifies most non-abusive instances in this class and rarely misses one.

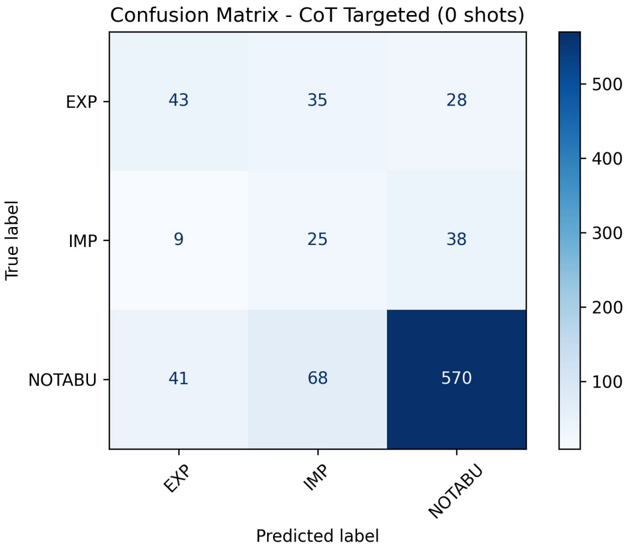


Figure 6.1: Confusion matrix for Qwen2.5 using the CoT Targeted prompting strategy (0-shot).

The performance on the EXP class is also moderate, with an F1-score of 0.43. Both the precision (0.46) and recall (0.41) suggest that the model makes a fair number of correct predictions with a fair number of true cases. However, the moderate scores still highlight the need for improvement in distinguishing explicit abuse from other categories.

The IMP class shows the weakest performance, with an F1-score of 0.25. There is also a noticeable imbalance between recall (0.35) and precision (0.20). This means that while the model captures some implicit abuse cases, many predictions are incorrect. This indicates that the model often misclassifies non-abusive or explicit instances as implicit abuse.

The classification report shows room for improvement, particularly in the EXP class, and even more so in the IMP class, where the models mainly struggle to grasp correct implicit instances.

6.1.2 Confusion Matrix

To better understand where the model exactly confuses labels, I examined the confusion matrix between classes, illustrated in Figure 6.1.

The best-performing class, NOTABU, was correctly classified by the model in 570 out of 682 instances (83.6%). This indicates a strong capability of the system to identify non-abusive content. However, there is still notable confusion: 41 instances (6.0%) were misclassified as EXP and 68 instances (10.0%) as IMP.

For the EXP class, 43 out of 106 instances (40.5%) were correctly classified, while 35 instances (33.0%) were incorrectly identified as IMP. Additionally, 28 instances (26.4%) were misclassified as NOTABU. This shows that the model struggles to clearly distinguish between explicit and implicit abuse, as well as between explicit abuse and non-abusive content.

For the IMP class, 25 out of 72 instances (34.7%) were correctly classified, while 9 instances (12.5%) were misclassified as EXP, and 38 instances (52.8%) as NOTABU, making this the largest subset of confusion. This indicates that the model particularly struggles with distinguishing implicit from non-abusive content.

The confusion matrix shows that the greatest room for improvement for the model lies in (1) differentiating IMP from NOTABU within the IMP class, followed by (2) distinguishing EXP from IMP within the EXP class.

6.2 Linguistic patterns in IMP confusions

6.2.1 Approach

To better understand the confusion made within IMP (true) as NOTABU (predicted) and EXP (true) as IMP (predicted), I manually examined these cases to identify any consistent linguistic patterns that may have led to the misclassification.

What we want to understand is where the system fails to detect the linguistic characteristics associated with explicit, implicit, and non-abusive speech. As an approach, I started from my own checklist of linguistic features for the three classes (of which further details are outlined in subsection 4.2.1). For clarity, I repeat the table here:

Explicit Abuse	**Implicit Abuse**	**Not Abusive**
Profanity	Limited or no profanity	Politeness
Direct pronouns and names	Stereotyping and generalizations	Empathy
Uppercase words and exclamations	Rhetorical questions	Neutral or positive tone
Syntactic simplicity	Complex or longer words	Syntactic complexity
Imperative structure	Coded or euphemistic language	Constructive critique
Direct insults, incl. slurs	Sarcasm and irony	
	False politeness / passive-aggressiveness	

Table 6.2: Linguistic features across abuse types.

Other approaches of checking in which cases an abusive speech model tends to fail include the set of functionalities for hate speech detection identified by HateCheck (Röttger et al., 2021). This checklist focuses on specific linguistic capabilities a model needs to have in order to demonstrate an understanding of hate speech. I reviewed this list to see if there were any additional functionalities that could supplement my checklist and selected those relevant to implicit abuse. These functionalities target more fine-grained linguistic phrasings rather than broader patterns and are intended for use in functionality tests. I used them as a substitute for my own checklist to explore whether they could help identify more specific patterns of implicit phrasing that the system needs to detect. This selection includes:

- **Normative threats** – Abuse masked as moral/legal opinion ("should be punished").
- **Abuse via negated positives** – "They are not smart" (actually negative).
- **Non-abuse via negated abuse** – "They're not dumb" (positive meaning).
- **Abuse phrased as a question** – e.g., "Why are they so violent?"
- **Abuse as opinion** – Framed with hedges like "I think that. . . "

In the next subsections, I begin with general observations on the detection of implicit abuse in tweets. I then present my findings on the two largest subsets of errors, IMP labeled as NOTABU and EXP labeled as IMP, using both my checklist and the partial HateCheck list introduced above as a framework. For each subset, I go over the most important linguistic patterns in the data to identify where the system fails to classify the content correctly. A full overview of the tweets, including my annotations for each subset, can be found in Appendix 8.

6.2.2 Understanding abuse in tweets

At the start of the error analysis, I observed that some tweets were difficult to understand, both in their literal content and implied meaning.

This is, firstly, because it can be hard to grasp the implied meaning when dealing with tweets. Implied meaning is often conveyed through an undertone or by referring to a broader context, which is easier to interpret in full conversations. Because tweets are short text, often viewed outside a clear context, the task becomes more difficult.

Secondly, the typical characteristics of tweets make them even harder to read: the use of in-group hashtags (e.g. #MAGA, referring to Donald Trump's "Make America Great Again" slogan, or #bb20, referring to season 20 of the reality show Big Brother); informal language, including spelling errors or ambiguous symbols. The following tweet illustrates the potential complexity of the linguistic characteristics found in this dataset:

> "#Liberals Are Reaching Peak Desperation To Call On #PhillipRuddock To Talk With #Turnbull To Convince Him To Help with #WentworthVotes 18 Sept 2018 @USER #Auspol #LNP #NSWpol @USER @USER @USER #LNPMemes URL" (ID 80397)

To fully understand this tweet, you need to be familiar with the political context, know who Phillip Ruddock is, interpret the informal language, and understand the meaning of the hashtags. These features make it challenging, even for the reader, to understand the implied meaning.

6.2.3 IMP (true) as NOTABU (predicted)

The example above is one of the tweets from the first subset: implicit abusive speech that was mislabeled as non-abusive by the model. This type of error occurred in 38 instances in the test set. To better understand what caused the errors in this subset, I systematically went through my checklist and the adopted HateCheck list for non-abusive and implicit speech.

The challenge of detecting characteristics within implicit abusive speech is that it is often masked as non-abusive. As shown in Table 6.2, features of implicit speech such as syntactic complexity, false politeness or empathy, and abuse framed as constructive critique can appear similar to those in non-abusive speech. Although the tweets in the confusion between IMP and NOTABU show features typical of non-abusive speech, none of them are truly non-abusive, because all of these cases carry an implied meaning of abuse.

To examine where these tweets show characteristics of non-abusive language, I first discuss overlapping features of the two classes found in the data, then features shared by implicit and non-abusive speech, and finally, features unique to only one of the two classes.

Overlapping features

I detected two patterns of features in the subset in which the non-abusive and implicit classes overlap.

Limited or no profanity — The first check I conducted focused on the presence of profanity, a characteristic in both non-abusive and implicit abusive speech.

This check showed that the tweets in this confusion contain little to no profanity; the few instances that did include it used relatively mild forms. This aligns with what is typically expected in non-abusive speech. In total, I identified only the following six occurrences:

- "f*cking" (ID 68875)
- "sh*t" (ID 97610, 81890, 91430, 91472)
- "ass" (ID 71350)

Complex or longer words — The second feature involved the length of the tweets. At first glance, the sentences in this subset appear longer, suggesting the presence of longer words or greater syntactic complexity. This can also be a characteristic of both implicit and non-abusive speech.

To verify this, I compared the average tweet length in this subset to that of the full dataset. On average, tweets in this subset are 10.13 characters longer (156.29 vs. 146.16), confirming that they tend to be longer. This is consistent with features of both non-abusive and implicit abusive speech.

Similar features

In addition to exactly overlapping features, the tweets also shared several similar characteristics, with the key difference lying in their implied meaning.

(False) politeness sarcasm or irony – A prominent pattern is the use of sarcasm and irony, often in the form of false politeness. This type of implicit abuse may seem polite on the surface, but it is actually used to express mockery or criticism. I found seven such examples. A few examples of this are:

- "Awwww she is so stinking cute! How old is she now?" (ID 73642)
- "Thanks a lot, now I've paid a fortune (...)" (ID 91430)
- "Please explain how this makes America great again." (ID 60466)

The first example could also be completely seen as an empathic and positive compliment, but the context — which is also not really clear to the reader — makes it implied abuse. We see the use of words like "thanks" and "please", suggesting a form of politeness, while the opposite is intended.

Empathy / rhetorical questions — Another frequent pattern is the use of fake empathy in the form of rhetorical questions. This feature also overlaps with the "Abuse phrased as a question" functionality in HateCheck. This way of phrasing is often used to imply criticism or to question someone's intelligence or morality, with the intended meaning embedded in the question itself.

Six tweets (16%) include this type of implicit abuse. The following example illustrates this feature:

- “Awwww she is so stinking cute! How old is she now?” (ID 73642)
- “To demonstrate how wholly incapable of carrying a thought to its logical conclusion she is?” (ID 58287)

These examples are phrased as questions, but mockingly asking about someone’s age or questioning their capability conveys the implied abusive meaning.

Constructive opinions / Abuse masked as a moral or opinion — Non-abusive language can express critique or opinion, but always in a constructive way.

I found three tweets that state an opinion in a polite or seemingly neutral way, but are implicitly abusive. These three examples are:

- “’Cause everyone knows a real leader falls into a van like a sack of potatoes...” (ID 79934)
- “If you want more death, listen to the ACLU, Antifa, Black Lives Matter” (ID 76833)
- “But here’s my opinion (...) I just think offset need to sit his ugly ass down.” (ID 71350)

Distinct features

Lastly, I found features that appear only in the implicit class. No features were found that appeared exclusively in the non-abusive class.

Stereotyping and Generalizations — The most frequently occurring pattern of implicit abuse in this subset is the use of stereotyping and generalizations, primarily targeting the political group liberals. This also highlights the overrepresentation of political content in the dataset.

In total, 10 tweets contain abusive generalizations in which liberals are portrayed through negative stereotypes. Examples of such generalizing tweets include:

- “#Liberals try this EVERY time (...)” (88490)
- “liberals probably don’t realize this because they are children (...)” (15815)
- “Liberals’ Favorite Myth (...)” (43782)

Furthermore, this linguistic pattern also includes two tweets that generalize groups based on national or ethnic origin, and one that targets women by mocking the #BelieveAllWomen movement:

- “When Liberals ask why you’re against illegal immigration? Taco Bell Employee: No Habla Ingles!” (ID 3129)
- “Bundesliga teams are competitive in euro competitions. Mexico hasn’t produced any talent like Pulisic in a long time (...)” (ID 97610)
- “#BelieveAllWomen just when I think things can’t get dumber (...)” (ID 15815)

This brings the total to 13 tweets (32%) in this subset that rely on stereotyping and generalizations.

Coded language — The last detected pattern is coded language. This is the most difficult to identify, as it is highly context-dependent and often relies on shared background knowledge. I found several examples that suggest this kind of implicit abuse, such as:

- "Money Soros" (ID 81890) — refers to a common antisemitic conspiracy theory
- "ex-crackhead" (ID 92215) — using moral or intellectual inferiority to insult a group
- "SJW bullies" (ID 57326) — refers to "Social Justice Warrior" in an insulting way
- "Antifa" (IDs 12193, 76833, 23542, 79222, 76135) — short for "anti-fascist", often used as a coded political signal

Findings

Table 6.3 illustrates a summary of my findings. Overall, the subset shows strong overlapping and similar features between the two classes. Only two features are solely characteristic of the implicit class; no features exclusive to the non-abusive class were present in the tweets.

Overlapping features	**Findings**
Limited or no profanity	6 cases (15.8%) — mild forms only
Complex or longer words	+10.13 characters above overall avg.
Similar features	**Findings**
(False) politeness / Sarcasm or irony	7 cases (18.4%)
Empathy / Rhetorical questions	6 cases (15.8%)
Constructive opinions	3 cases (7.9%)
Distinct features (only in IMP)	**Findings**
Stereotyping and generalizations	13 cases (34.2%)
Coded or euphemistic language	8 cases (possibly more) (21.1%)

Table 6.3: Findings on linguistic features in tweets labeled as implicit (IMP) but predicted as not abusive (NOTABU).

So why did the system possibly fail to classify an example like the one shown at the beginning of this subsection (ID 80397, see subsection 6.2.2)? The issue in these cases is that the lack of profanity and the presence of longer sentences overlap with characteristics of the NOTABU class. In addition, false politeness can easily be confused with genuine politeness and empathy, and abuse phrased as an opinion or a question can be mistaken for constructive critique or empathy. Stereotyping, generalizations, and coded language do appear frequently in the data, but the system was not able to grasp their meaning when embedded in seemingly polite or neutral phrases. When the political context is unknown, "#Liberals Are Reaching Peak Desperation To Call On #PhillipRuddock To Talk With #Turnbull To Convince Him To Help with #WentworthVotes" may just seem like a neutral statement to the system.

To detect implicit abuse, the system cannot rely on these surface features alone and must be able to capture undertones such as sarcasm, rhetorical questions, or coded language, usually requiring context. This overlap in surface features may be a likely reason the system fails to detect implied meanings and labels the data as non-abusive.

6.2.4 EXP (true) as IMP (predicted)

The second largest subset of misclassifications is the confusion of EXP as IMP, consisting of 35 tweets. This group includes cases of explicit abusive speech that were labeled as implicit abuse. Below, I will list the most occurring linguistic patterns of differences of explicit and implicit speech present in the data.

Level of explicitness

At first glance, the difference between this subset (EXP as IMP) and the previous one (IMP as NOTABU) is immediately noticeable. The general tone is more aggressive, direct, and easier to interpret. To determine the level of explicitness in the text, I identified several patterns.

Profanity — First, I examined the use of profanity in this subset of tweets. The level of profanity is a distinguishing feature between the two classes. Although implicit language can include profanity, it is typically milder, whereas the explicit class tends to use it in a more direct and aggressive manner.

In total, 9 tweets contain some form of profanity. On average, the level of aggression remains relatively mild to moderate. The profanity ranges from milder forms—such as "sh*t" (IDs 46444, 55633, 45269), "b*tch" (ID 70569), and "f*cking" (IDs 46444, 32061)—to more severe variations, such as:

- "Holy sht, please your fcking *ssholes. (...)" (ID 46444)
- "(...) TWITTERFUCK *SS" (ID 75125)

Simplicity of wording — Implicit speech often shows more syntactic complexity. To check whether sentence length correlates with these errors, I calculated the average tweet length. This subset of errors has a similar tweet length to the overall dataset (147.06 vs. 146.16), indicating that there is no clear relationship between tweet length and this type of confusion.

However, we do observe some degree of syntactic simplicity in this subset. In particular, there are five clear cases of imperative structures, including:

- "Stop talking about everyone else and start talking (...)" (ID 67049)
- "Stop the fascist NVU in #Amsterdam!" (ID 70443))

Direct insults and slurs / Direct pronouns and named references — Where implicit speech is often expressed in more subtle ways, explicit abuse often includes direct insults and the use of personal pronouns or named individuals to make the target of the abuse clear. This is one of the most common patterns in the subset, of which I identified 16 cases. These cases show clear, straightforward insults. Examples from this pattern include:

- "SierraBurgessIsALoser (...)" (ID 24049)
- "she is fat ugly libreal" (ID 67841)

- “I hope he rots in hell!” (ID 22067)

Uppercase words and exclamations — Another often occurring pattern is the use of uppercase words in 6 instances, used to emphasize anger or intensity. Additionally, 12 tweets include exclamation marks. Examples include:

- “SO WANT TO KICK TWITTERF*CK A*S.” (ID 75125)
- “(...) EVERYTHING LIBERALS TOUCH TURNS TO ABSOLUTE SH*T.” (ID 45269)
- “#Carrey is a #pervert himself!” (ID 79778)

The data shows that the level of explicitness in this subset ranges from mild to moderate in intensity. Profanity is present but not extreme. More severe patterns appear through simple wording, direct insults, and the use of uppercase letters.

Implicit features

Although the tweets in this subset are labeled as explicit and show clear features of explicitness, the system still misclassified them as implicit. This suggests that the data also contains patterns commonly associated with implicit abuse. To better understand this, I identified three linguistic patterns of implicitness that may have influenced the model’s predictions.

Rhetorical questions — Abuse framed as a question is a common trait of implicit forms of abusive speech (as also discussed in Subsection 6.2.2), but it can also occur in explicit forms of abuse.

In the data, I identified four rhetorical questions (11.4%). The following two examples were annotated as explicit but predicted as implicit:

- “Wtf does she think she is, a person? Pffft pathetic” (ID 21524)
- “And she has a pet??? F*cking disgusting” (ID 32061)

What stands out in these two cases is the use of direct and explicit language. The tweets refer to the target as “pathetic” or “f*cking disgusting”. While rhetorical questions are often associated with implicit abuse, in these examples, the abusive intent is clearly explicit due to the added direct wording. The model does not appear to fully capture these subtle distinctions between explicit and implicit abuse.

Embedded hashtags — In the data, we see that some tweets express abuse through embedded hashtags, where the abusive content is contained within the hashtag itself, changing the sentiment of the overall tweet. These cases are difficult to count precisely, as their interpretation is highly context-dependent, and the hashtags can be challenging to read, for both humans and the system. There were 9 examples (25.7%) of these cases present in the data. A few examples include:

- “SierraBurgessIsALoser” (ID 24049)
- “LiberalHypocrisy” (ID 76379)
- “DemocratPartyOfDomesticTerrorists” (ID 52080)

Abuse masked as moral or opinion — Another pattern that occurs in the data is abuse framed as personal or moral opinion. In these cases, the abusive language is embedded within what appears to be a personal or moral view. This phrasing makes it more difficult for the model to recognize the abuse as explicit, as the aggression is presented as reasoning or belief rather than a directly targeted insult. In total, I found five such cases (15.6%). The following examples display this pattern:

- "Holy shit, please your fucking assholes, don't blame someone for the death of other one. She is sad enough for today, don't you see? It isn't fault of none, he had an overdose and died. End. Stop wanting someone to blame, fuckers." (ID 46444)

- "Conservatives can't debate honestly, and they have no integrity... They're fundamentally dishonest people." (ID 46139)

Findings

Table 6.4 illustrates a summary of my findings. Overall, the subset showed clear characteristics of explicit abuse, but with, on average, only a mild to moderate presence of profanity and slurs. Most of the abuse was expressed through direct insults and personal pronouns, sometimes combined with exclamations, occasional use of uppercase words, and a few imperative structures.

Level of explicitness	**Findings**
Profanity	9 cases (25.7%)
Direct insults and slurs / Direct pronouns and named references	16 cases (45.7%)
Uppercase words and exclamations	18 cases (51.4%)
Syntactic simplicity / Imperative structures	+0.90 tokens from overall avg., 5 cases (14.3%)
Implicit features	**Findings**
Rethorical questions	4 cases (11.4%)
Embedded hashtags	9 cases (25.7 %)
Abuse masked as moral or opinion	5 cases (15.6%)

Table 6.4: Findings on linguistic features in EXP as IMP.

However, the error subset also contains forms of implicit abuse, such as rhetorical questions, embedded hashtags, and abuse masked as moral or personal opinion. The distinction between IMP and EXP often hinges on subtle cues, such as explicitly naming the target, as seen in the reference to "a person" (ID 21524). Hashtags that embed abusive meaning further complicate this distinction. A tweet can contain explicit abuse while still being phrased as a question or framed as a personal belief, and in some cases, the sentiment can shift easily with the use of a hashtag.

Overall, the system does not yet fully grasp the boundary between implicit and explicit abuse when explicit language is embedded in rhetorical or more indirect forms. To better identify this distinction, the model needs to recognize that abusive profanity carries more weight than the rhetorical structure in which it appears. Detecting direct references to a target or understanding the meaning of a hashtag also seems to be a key factor in distinguishing explicit abuse from implicit forms.

6.3 Automated Analysis

In the automated analysis, I conducted three experiments on subsets that are expected to be challenging for the model. First, I examined the labels where the annotations do not overlap between the OLID/OffensEval and AbuseEval datasets. As there are mismatches between the annotation guidelines of offensive and abusive, it is interesting to see whether these borderline cases also end up being more confusing for the model. Second, I analyzed the overlap in correctly identified IMP labels across three different prompt strategies, to see if there are any cases the model can consistently identify. Lastly, I examined the effect of targetness on the classification errors using the OLID/OffensEval labels. The sections below explain all three in further detail.

6.3.1 Reannotation of OLID/OffensEval to AbuseEval

The reannotation of the OLID dataset to AbuseEval for EXP, IMP, and NOTABU also included a shift from offensive language to abusive language. This resulted in 62 cases in the test set that were originally labeled as offensive but are now annotated as non-abusive. The other way around, tweets that were not labeled as offensive but are now labeled as explicit or implicit, did not occur, because those tweets are not taken into consideration. As it is already difficult for humans to distinguish the differences between offensive and abusive, we can expect the system to struggle with these borderline cases as well. The discussion around disagreement in definitions is further outlined in subsection 2.1.1.

Table 6.5 shows the classification report for this subset, while Figure 6.2 displays the confusion matrix. Since we are looking at a subset consisting only of NOTABU examples, precision is not reliable. However, recall shows a drop from 0.84 of the NOTABU class on the full dataset to 0.52 in this subset. The model only correctly identified 32 (51.61%) of the NOTABU instances. 15 cases (24.19%) were incorrectly classified as explicit, and another 15 (24.19%) as implicit. This marks a significant drop in performance compared to the overall results on the NOTABU class.

Label	Precision	Recall	F1-score	Support
NOTABU	1.00	0.52	0.68	62

Table 6.5: Classification report on the mismatches in reannotation between OLID/OffensEval and AbuseEval.

The results suggest that the reannotation process indeed introduced confusion about what qualifies as abusive speech. A closer look at the data shows that misclassifications from NOTABU to EXP or IMP often involve profanity or slurs. Among these false positives, I observed the use of words like “n#gga(s)” (5 instances) or “b#tch(es)” (3 instances). The tone of these messages appears ambivalent; most seem intended in an amicable or conversational manner. In the AbuseEval dataset, not all instances of words often seen as insulting, such as “n#gga(s),” are annotated as offensive. This is likely chosen to prevent bias against groups that more frequently use these words in a friendly manner. It is debatable whether these kinds of instances constitute offensive or abusive speech, as the words carry inherently negative connotations but are used in a different, often non-hostile way, and are not targeted at a person or group. This highlights the difficulty and ambiguity involved in annotating and detecting abusive speech.

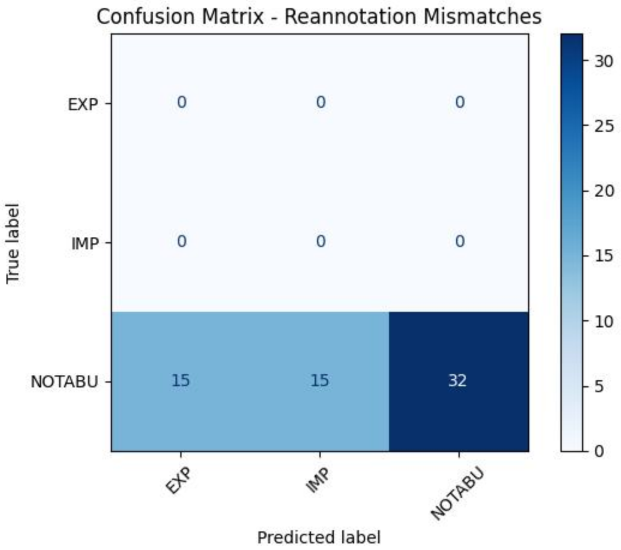


Figure 6.2: Confusion matrix on the mismatches in reannotation between OLID/OffensEval and AbuseEval.

We also see that some tweets fall under borderline cases, such as "...What the fck did he do this time?" (ID 83681), which was misclassified by the system as IMP. The tweet uses the word "fck" and judges someone's behavior, but unlike the previous examples, it does not directly insult the person's identity. This is a very important difference that models need to be able to understand.

6.3.2 Overlap in IMP labels

In addition to analyzing the false predictions, it is also interesting to examine where the model is able to correctly classify instances. Several prompting strategies showed a similar number of correctly labeled implicit tweets. It is interesting to look at whether tweets are correctly identified due to the prompting strategy or because of certain inherent characteristics of the tweets themselves.

For this analysis, I compared the predictions from three strategies: the best-performing model I analyzed in the manual part—CoT Targeted (0-shot), Base (8-shot), and CoT IHC Labels (0-shot). All three show similar performance on implicit abuse. The full results are shown in Table 5.5. CoT Targeted (0-shot) correctly identified 25 implicit examples, Base (8-shot) identified 24, and CoT IHC Labels (0-shot) identified 30.

Table 6.6 shows the counts among the different strategies. The overlap between any two strategies ranges from 10 to 14 labels. There is an overlap of only 6 cases across the three strategies. These results, given the small overlap of correctly identified cases across all three, suggest that it is primarily the prompting strategy, rather than the intrinsic characteristics of the tweets, that results in correct identification.

However, when looking closer at the six cases identified by all strategies, some common characteristics do emerge. We see that these examples show clearer, directer signs of abusive speech, likely making it easier for the system to detect them. We see several slurs, such as "n#gga(s)" (ID 97410), and degrading comparisons like "#Conservatives just like old #Garbage bags" (ID 46983). There is also demeaning language, for ex-

Metric	Count
IMP total	
CoT targeted (0 shots)	25
Base (8 shots)	24
CoT IHC labels (0 shots)	30
IMP in common	
Targeted vs IHC labels	14
Targeted vs Base	10
IHC labels vs Base	11
All three	6
Unique IMP	
CoT targeted (0 shots)	7
Base (8 shots)	9
CoT IHC labels (0 shots)	11

Table 6.6: Summary of IMP predictions across three prompting strategies.

ample in the tweet: "a grown ass woman, probably 10 years older than me is currently spreading rumors rather than talking to me about it, nice work you got there satan" (ID 33394).

These tweets, while implicit, are not subtle. They are clearly harmful, mostly target individuals or groups, and use abusive words that don't require much contextual understanding. This shows that subtle cases of implicit abuse, such as implied meanings in rhetorical questions and sarcasm that are not present in the true positives, are still hard to detect for the three prompt strategies.

6.3.3 Effect of targetness

In the manual analysis and earlier sections about related work (chapter 2), we have seen that targetness appears to be a key factor in understand abusive speech. Therefore, I lastly examined whether there are any meaningful differences in the distribution of targetness across the results, using the labels from the OLID/OffensEval annotation (as described in subsection 3.1.1).

Since these are labels from the OLID/OffensEval annotation, this analysis only applies to instances that were classified as offensive in the original annotation and abusive in the reannotation. For this, I compared the percentages of correctly and incorrectly identified cases across the different classes, since the distribution of TIN and UNT is skewed.

As shown in Figure 6.3, the ratio of targeted (TIN) to untargeted (UNT) instances remains roughly the same, regardless of whether the prediction was correct or not. This consistent pattern suggests that targetness does not play a role in explaining specific types of classification errors.

6.3.4 Summary

The error analysis offers a deeper understanding of when the model classifies and misclassifies abusive content.

The overall performance of the model is moderate, with a macro-average F1-score of 0.52. While the system correctly identifies most non-abusive instances in the NOTABU class, it mainly struggles to grasp correct implicit instances. The confusion matrix

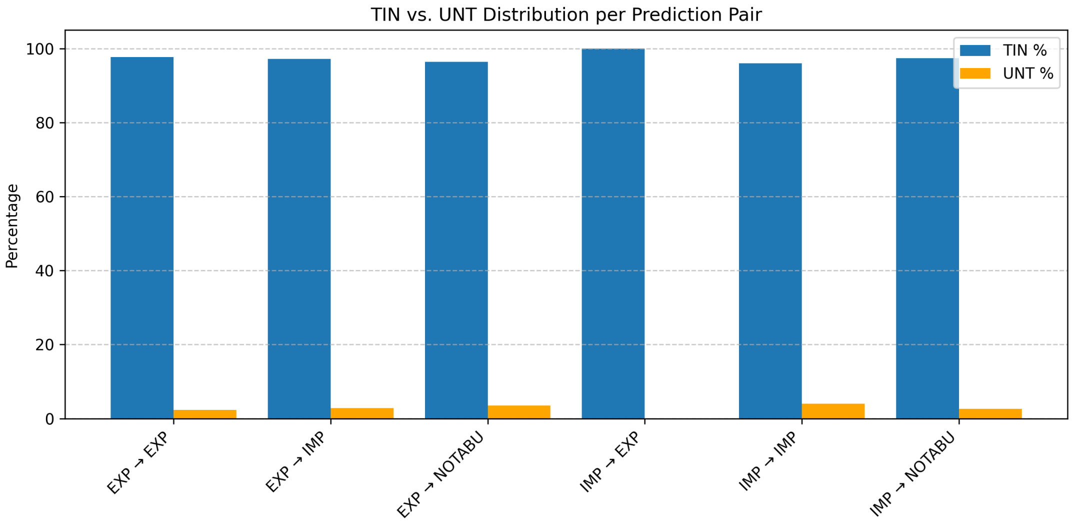


Figure 6.3: Distribution of targetness per confusion (TIN = targeted, UNT = untargeted).

shows that the greatest room for improvement for the model lies in differentiating IMP from NOTABU within the IMP class, followed by distinguishing EXP from IMP within the EXP class.

A first glance at the errors shows that understanding abuse in tweets is inherently challenging due to the lack of context, informal and messy language, and the need for background knowledge to grasp the implied meaning.

In the IMP as NOTABU errors, tweets contained little to no profanity, but showed the use of stereotyping and generalizations, often embedded in rhetorical questions, sarcasm and irony, or coded or euphemistic language. These tweets were longer, which is in line with the features of implicit abuse. However, false politeness can be easily confused with real politeness and empathy, and abuse phrased as a question or opinion can be confused with constructive critique. To detect implicit abuse, the system cannot rely on these features alone and must capture undertones such as sarcasm, rhetorical questions, or coded language, usually relying on context. This overlap may be a likely reason the system fails to detect implied meanings and labels the data as non-abusive.

In the EXP as IMP errors, it became clear that the system does not yet fully grasp the boundary between implicit and explicit abuse when explicit language is embedded in rhetorical or more indirect forms. To better identify this distinction, the model needs to recognize that abusive profanity carries more weight than the rhetorical structure in which it appears. Detecting direct references to a target or understanding the meaning of a hashtag also seems to be a key factor in distinguishing explicit abuse from implicit forms.

The small differences the system has to handle are best shown in the following examples from subsets IMP as NOTABU, EXP as IMP, and the reannotation mismatches from the error analysis:

- “Wtf does she think she is, a person? Pffft pathetic” (ID 21524) — labeled as EXP, predicted as IMP
- “Who the hell does he think he is?” (ID 89200) — labeled as IMP, predicted as

NOTABU

- “What the fuck did he do this time?” (ID 83681) — labeled as NOTABU, predicted as IMP

These examples show how subtle the differences between the classes can be, and that the model often struggles to determine where to draw the line between explicit, implicit, and non-abusive language. The first sentence directly insults the person’s identity. The second also targets a person’s identity, but is phrased as a rhetorical question. The third example uses profanity but attacks only someone’s actions, not their identity. Together, these examples illustrate the importance of the model understanding the line where targetness starts.

The automated part of the analysis showed that reannotation mismatches indeed introduce more confusion about what qualifies as abusive speech, especially in borderline cases that include profanity but are used in a different, often non-hostile way. It is debatable whether instances such as “n#gga” constitute offensive or abusive speech, as the words carry inherently negative connotations but are used in a different, often non-hostile way and are not targeted at a person or group. This highlights the difficulty and ambiguity involved in annotating and detecting abusive speech.

When examining correctly identified IMP cases across prompt strategies, there is an overlap of only 6. These tweets were not subtle, clearly harmful, and used abusive words that do not require much contextual understanding. This suggests that more subtle cases of implicit abuse, such as implied meanings in rhetorical questions and sarcasm, are still hard to detect. Moreover, targetness does not play a role in explaining specific types of classification errors, but is equally present in all subsets of confusion and correct identifications.

Chapter 7

Discussion

This study answers the question of whether prompt engineering a generative model offers advantages over a fine-tuned BERT model, particularly in identifying implicit cases of abusive speech. The results showed that, although Qwen demonstrates a solid understanding of what constitutes abusive language, the fine-tuned BERT model still outperforms the prompted Qwen in both explicit and implicit cases. The following sections provide a detailed analysis to support these findings.

Fine-tuning BERT

Firstly, I explored the performance of a fine-tuned BERT on the AbuseEval dataset. The setup of the experiments is outlined in section 4.1.

The results showed that the binary and ternary classification tasks are differently affected by the choice of model and training data. For binary classification, HateBERT trained solely on the AbuseEval dataset achieved the highest performance, suggesting that (although minimal) domain-specific pretraining can be beneficial. In contrast, BERT trained on the combined dataset showed the best performance for ternary classification. While the scores for implicit cases remain lower than for other categories, this category shows the most notable improvement when using the combined dataset.

The difference in binary and ternary classification performance on HateBERT suggests that the model relies more on explicit features, likely due to its training on more overtly abusive content. This highlights both the challenges of detecting implicit abuse and the value of more varied training data to help the model generalize to such cases. The traditional model BERT trained on the combined training set served as a baseline for the following prompting experiments.

Prompting Qwen

Secondly, I explored which prompting strategies are most effective in improving the detection of abusive speech. The setup of the experiments is outlined in section 4.2.

Across both binary and ternary classification setups, incorporating CoT or CoT with targetness proved most effective in improving prompt performance. The best-performing models were CoT (1-shot, temperature = 0.8) for binary classification and CoT Tar (0-shot, temperature = 0.2) for ternary classification. Changes in temperature showed only a minimal effect on the scores.

A total number of 16 shots was defined as the ”sweet spot” in performance in previous work by Han and Tang (2022). The binary results do not support that 8 shots (16 in total) are significantly better than 1 or even 0 shots, and there is no consistency in the 8-shot experiments. For the ternary task, the scores for 6 shots (18 in total) are good, but also not consistently better than other numbers of shots. The results show that the prompting strategy, in this case, has more impact on the model's performance than the exact number of prompts.

Further prior research showed that adding informative instruction and requesting a CoT are among the most effective prompting techniques. Other research also showed that including target information or explanations is beneficial for the model's performance (He et al., 2021). Across both binary and ternary classification setups, incorporating CoT or CoT with targetness is most effective in improving prompt performance, aligning with findings from previous work. In addition, instructing the model to consider the IHC labels within the CoT also led to improvements in detecting implicit cases.

This shows that the model performs particularly well when its own knowledge is activated through reasoning, and that adding a (guided) reasoning step is beneficial for the detection of abusive speech. It also suggests that, when tested on the AbuseEval dataset, the model already has a solid understanding of what implicit abuse, explicit abuse, and non-abusive language look like.

Error Analysis

Finally, I conducted a thorough error analysis on the best-performing prompting strategy, CoT with targetness (0-shot), to answer which patterns in the errors, and in particular implicit abusive speech, affected the model's results. The overall outline and results are presented in chapter 6.

The overall performance of the model was moderate, with a macro-average F1-score of 0.52. While the system correctly identified most non-abusive instances in the NOTABU class, the greatest room for improvement lies in cases where IMP is confused with NOTABU and EXP with IMP. The characteristics of tweets contribute to the challenge: understanding abuse in tweets is inherently challenging due to the lack of context, informal and messy language, and the need for understanding of context to grasp the implied meaning.

In the IMP as NOTABU errors, I observed overlapping characteristics between the two classes. This overlap was likely a reason the system failed to detect implied meanings and labeled the data as non-abusive. In the EXP as IMP errors, it became clear that the system did not yet fully grasp the boundary between implicit and explicit abuse, particularly when explicit language appeared in indirect or rhetorical forms. Recognizing that abusive profanity often carries more weight than rhetorical structure, and detecting abuse in the form of targetness or hashtags, are key challenges for the model.

The following examples illustrate how subtle the boundaries between the three classes can be, and that the model often struggles to determine where to draw the line between explicit, implicit, and non-abusive language:

- “Wtf does she think she is, a person? Pffft pathetic” — EXP, predicted IMP
- “Who the hell does he think he is?” — IMP, predicted NOTABU

- "What the fuck did he do this time?" — NOTABU, predicted IMP

The first sentence directly insults the person's identity. The second also targets the identity of a person, but is phrased as a rhetorical question. The third example uses profanity but attacks only someone's actions, not their identity. Together, these examples illustrate the importance of the model understanding the line where targetness starts.

As earlier research had pointed out (discussed in chapter 2), there is no universal definition of what is considered abusive speech, which creates inconsistencies in annotation guidelines across researchers. The error analysis underlines that reannotation mismatches between OffensEval/OLID and AbuseEval indeed introduced more confusion, especially in borderline cases with profanity used in a conversational manner (such as n#gga in certain communities). This highlights the difficulty and ambiguity involved in annotating and detecting implicit abuse.

The overlap of only 6 correctly identified IMP cases across 3 strategies showed that different prompts relied on different features. This indicated that the model was only able to rely on more obvious signals, while implicit cases remained difficult to detect.

Overall, the error analysis showed that implicit cases affected the performance of the model due to errors caused by overlap in features with the non-abusive class and the system's difficulty in relying on explicit features when rhetorical structures, often seen in implicit cases, were used. The line between the three classes lay partly in the model's ability to understand when a text was targeted at someone or to understand the context, such as that implied by hashtags. The lack of universal consensus on what is considered abusive contributed to these challenges, and the system still seems to rely on more obvious forms of abuse when correctly classifying implicit cases.

Key findings

Despite some promising results in prompting, especially with CoT strategies, fine-tuning delivered better overall performance and outperformed all prompting approaches.

The traditional BERT model trained on the combined training set achieved the highest overall performance. The prompted Qwen model performed particularly well when its internal knowledge was activated through reasoning. However, implicit abusive speech negatively affects the model's performance, contributing to a substantial number of misclassifications. This was largely due to overlapping features with non-abusive content and the system's struggle in detecting where explicit or implied meaning begins.

The error analysis showed that improvements lie in helping the model to understand when a text is targeted at someone or to understand the broader context the abuse refers to. The lack of universal consensus on what is considered abusive contributed to these challenges.

Chapter 8

Conclusion

In this research, I answered the question of whether prompt engineering offers advantages over a fine-tuned BERT model, particularly in identifying implicit cases of abusive speech. The results showed that, although Qwen demonstrates a solid understanding of what constitutes abusive language, the fine-tuned BERT model still outperforms the prompted Qwen in both explicit and implicit cases.

Prior work showed difficulties in annotation guidelines and in detecting implicit cases of abusive speech. It also demonstrated promising results in detecting abusive speech in general and particularly implicit cases, using transformer-based models and prompting generative large language models (LLMs).

The fine-tuning experiments showed that the binary and ternary classification tasks are differently affected by the choice of model (BERT or HateBERT) and training data (solely AbuseEval, or AbuseEval and IHC), likely due to the model's inability to generalize to implicit cases when pre-trained on abusive speech. The BERT model trained on the combined AbuseEval and IHC dataset achieved the highest performance in ternary classification.

The results of the prompting experiments showed that across both binary and ternary classification setups, incorporating CoT or CoT with targetness proved most effective in improving prompt performance. This shows that the model performs particularly well when its own knowledge is activated through reasoning, and that adding a (guided) reasoning step is beneficial for the detection of abusive speech.

The error analysis showed that the model struggled with confusion between the implicit class and other classes, and that improvements lie in helping the model to understand when a text is targeted at someone or to understand the broader context the abuse refers to. The lack of universal consensus on what is considered abusive contributed to these challenges.

Limitations and future work

Limitations in this research are mainly due to the time and scope of the conducted experiments. It should be noted that the comparison in this study involved a relatively small fine-tuned model. Future work could investigate whether larger or more state-of-the-art fine-tuned LLMs outperform prompt-based approaches even more.

Additionally, this study focused on general BERT and the domain-specific HateBERT models, rather than models specifically trained on tweets. Testing a model type that focuses on these characteristics may offer potential improvements in the model's ability to understand tweets and their context.

Moreover, the AbuseEval dataset, which consists of political tweets, may have influenced model performance by relying on domain-specific patterns and limiting generalizability. Moreover, future work on prompting should also emphasize a more solid grounding in the definition of what constitutes abusive or offensive content within annotated datasets, and adjust prompting strategies accordingly so that systems are better capable of handling borderline cases.

Finally, the prompting setup was limited to configurations based on previous work. Prompting strategies that focus on reasoning, such as multiple reasoning paths through Tree of Thoughts (ToT), could potentially improve model performance. Further work focusing on improving prompts in terms of understanding targetness and context within the specific characteristics of tweets, and exploring how different types of targeted and untargeted examples the model was exposed to during the prompting experiments affected its performance, could provide valuable additional insights.

Appendix A

Prompt to ChatGPT on May 19, 2025 for subsection 2.1.2

User Prompt:

> Can you give an example of this? Only give the example, nothing else.
>
> Explicit and implicit abusive language have different linguistic features that are important to consider during the development of detection systems. In the case of explicit abusive speech, the language is unambiguous in its potential to be abusive: the literal definition of the words used according to the dictionary is abusive, such as language containing racist or homophobic slurs or profanity (Ocampo et al., 2023). It is more informal, angrier, and often explicitly attacks the target, making use of second-person pronouns or specific names, with fewer analytic words and more words suggesting authority and influence (Zampieri et al., 2019; ElSherief et al., 2018). The sentences are often short, fragmented, and grammatically simple, and imperative forms and aggressive syntactic structures are prevalent (Waseem and Hovy, 2016; Davidson et al., 2017).

ChatGPT Response (Explicit Example):

> *"You are such a worthless piece of trash. Go back to where you came from, freak."*

User Prompt:

> Can you do the same for this?
>
> Explicit slurs and profanity are overall absent in implicit abusive speech, which relies instead on more subtle and indirect forms of abuse. In an extensive study of hate speech datasets, Ocampo et al. (2023) identified 18 typical properties of implicitness based on linguistic characteristics. Among the most significant are irony, sarcasm, black humor, metaphor, exaggeration, rhetorical questions, sentiment, inference, lack of context, and absence of extralinguistic knowledge, often appearing in combination (Ocampo et al., 2023).

ChatGPT Response (Implicit Example):

> *"Oh great, another one of them got the job. Must be all that 'diversity' working its magic again."*

Appendix B

This appendix provides a full overview of the prompts used for the prompting experiments.

Prompts for Binary Classification

The examples used in the following prompts in binary set-up can be found in Table 1. The corresponding reasonings for CoT, CoT - Targeted, and CoT - IHC label strategies are listed in Table 2.

Strategy 1: Binary - Base

```
Classify the following texts into one of the following categories:
- abusive (ABU)
- not abusive (NOTABU)

Here are a few examples of labeled texts:
ABU: [Example abusive text 1]
ABU: [Example abusive text 2]
...
NOTABU: [Example non-abusive text 1]
NOTABU: [Example non-abusive text 2]
...

Now classify the following texts. Always choose ONE label per text.
Output ONLY a valid Python list of N labels. Do NOT explain. Do NOT add
anything else.

Texts:
1. [Text to classify 1]
2. [Text to classify 2]
...
N. [Text to classify N]
```

Strategy 2: Binary - Definition

```
Classify the following texts into one of the following categories:
- abusive (ABU): Language that is explicitly or implicitly offensive, hurtful
or profane. This includes speech that debases, insults, or expresses intense
negative emotion. This also includes subtle forms like sarcasm and passive-
```

```
aggression.
- not abusive (NOTABU): Language that is polite, respectful or neutral, and
free from explicit or implicit harmful intent.

Here are a few examples of labeled texts:
ABU: [Example abusive text 1]
ABU: [Example abusive text 2]
...
NOTABU: [Example non-abusive text 1]
NOTABU: [Example non-abusive text 2]
...

Now classify the following texts. Always choose ONE label per text.
Output ONLY a valid Python list of N labels. Do NOT explain. Do NOT add
anything else.

Texts:
1. [Text to classify 1]
2. [Text to classify 2]
...
N. [Text to classify N]
```

Strategy 3: Binary - CoT

```
Classify the following texts into one of the following categories:
- abusive (ABU)
- not abusive (NOTABU)

Instructions for each text:
1. Write 1{2 short sentences explaining why the text is classified as ABU or
NOTABU.
2. Then output the label on a separate line starting with: Label: ABU or
Label: NOTABU.

Follow this example format:
Reasoning: <reasoning>
Label: <ABU or NOTABU>

Here are a few examples of labeled texts:
Text: [Example abusive text 1]
Reasoning: [Reasoning for why it's abusive]
Label: ABU
...
Text: [Example non-abusive text 1]
Reasoning: [Reasoning for why it's not abusive]
Label: NOTABU
...

Now classify the following texts. ALWAYS output the instruction steps for EACH
```

```
of the N texts.

Texts:
1. [Text to classify 1]
2. [Text to classify 2]
...
N. [Text to classify N]
```

Strategy 4: Binary - Tar

```
Classify the following texts into one of the following categories:
- abusive (ABU)
- not abusive (NOTABU)

Instructions for each text:
1. Write 1{2 short sentences explaining why the text is classified as ABU or
NOTABU. Consider in your explanation if the text is targeted or untargeted.

2. Then output the label on a separate line starting with: Label: ABU or
Label: NOTABU.

Follow this example format:
Reasoning: <reasoning>
Label: <ABU or NOTABU>

Here are a few examples of labeled texts:
Text: [Example abusive text 1]
Reasoning: [Reasoning including whether it is targeted or untargeted]
Label: ABU
...
Text: [Example non-abusive text 1]
Reasoning: [Reasoning including whether it is targeted or untargeted]
Label: NOTABU
...

Now classify the following texts. ALWAYS output the instruction steps for EACH
of the N texts.

Texts:
1. [Text to classify 1]
2. [Text to classify 2]
...
N. [Text to classify N]
```

Strategy 5: Binary - IHC

```
Classify the following texts into one of the following categories:
- abusive (ABU)
- not abusive (NOTABU)

Instructions for each text:
1. Write 1{2 short sentences explaining why the text is classified as ABU or
NOTABU. Consider in your explanation if the text includes:
- white grievance
- incitement to violence
- inferiority language
- irony
- stereotypes and misinformation
- threatening and intimidation
2. Then output the label on a separate line starting with: Label: ABU or
Label: NOTABU.

Follow this example format:
Reasoning: <reasoning>
Label: <ABU or NOTABU>

Here are a few examples of labeled texts:
Text: [Example abusive text 1]
Reasoning: [Reasoning referencing applicable IHC cues]
Label: ABU
...
Text: [Example non-abusive text 1]
Reasoning: [Reasoning referencing applicable IHC cues]
Label: NOTABU
...

Now classify the following texts. ALWAYS output the instruction steps for EACH
of the N texts.

Texts:
1. [Text to classify 1]
2. [Text to classify 2]
...
N. [Text to classify N]
```

Prompts for Ternary Classification

The examples used in the ternary set-up can be found in Table 3. The corresponding reasonings for the CoT, CoT - Targeted, and CoT - IHC label strategies are listed in Table 4.

Strategy 1: Ternary - Base

```
Classify the following texts into one of the following categories:
- explicit abuse (EXP)
- implicit abuse (IMP)
- not abusive (NOTABU)

Here are a few examples of labeled texts:
EXP: [Example explicit abusive text 1]
IMP: [Example implicit abusive text 1]
NOTABU: [Example non-abusive text 1]
...

Now classify the following texts. Always choose ONE label per text.
Output ONLY a valid Python list of N labels. Do NOT explain. Do NOT add

anything else.

Texts:
1. [Text to classify 1]
2. [Text to classify 2]
...
N. [Text to classify N]
```

Strategy 2: Ternary - Definition

```
Classify the following texts into one of the following categories:
- explicit abuse (EXP): Language with direct and literal forms of abusive
speech, such as slurs, profanity, and other clearly hostile or offensive
expressions.
- implicit abuse (IMP): Language with subtle and indirect forms of abusive
speech, such as passive-aggressiveness, sarcasm with harmful undertone, and

other forms with implied harm.
- not abusive (NOTABU): Language that is polite, respectful or neutral, and

free from explicit or implicit harmful intent.

Here are a few examples of labeled texts:
EXP: [Example explicit abusive text 1]
IMP: [Example implicit abusive text 1]
NOTABU: [Example non-abusive text 1]
...

Now classify the following texts. Always choose ONE label per text.
Output ONLY a valid Python list of N labels. Do NOT explain. Do NOT add

anything else.
```

```
Texts:
1. [Text to classify 1]
2. [Text to classify 2]
...
N. [Text to classify N]
```

Strategy 3: Ternary - CoT

```
Classify the following texts into one of the following categories:
- explicit abuse (EXP)
- implicit abuse (IMP)
- not abusive (NOTABU)

Instructions for each text:
1. Write 1{2 short sentences explaining why the text is classified as EXP or
IMP or NOTABU.
2. Then output the label on a separate line starting with: Label: EXP or
Label: IMP or Label: NOTABU.

Follow this example format:
Reasoning: <reasoning>
Label: <EXP or IMP or NOTABU>

Here are a few examples of labeled texts:
Text: [Example explicit abusive text 1]
Reasoning: [Reasoning for EXP]
Label: EXP

Text: [Example implicit abusive text 1]
Reasoning: [Reasoning for IMP]
Label: IMP

Text: [Example non-abusive text 1]
Reasoning: [Reasoning for NOTABU]
Label: NOTABU
...

Now classify the following texts. ALWAYS output the instruction steps for EACH
of the N texts.

Texts:
1. [Text to classify 1]
2. [Text to classify 2]
...
N. [Text to classify N]
```

Strategy 4: Ternary - Tar

```
Classify the following texts into one of the following categories:
```

```
- explicit abuse (EXP)
- implicit abuse (IMP)
- not abusive (NOTABU)

Instructions for each text:
1. Write 1{2 short sentences explaining why the text is classified as EXP or
IMP or NOTABU. Consider in your explanation if the text is targeted or
untargeted.
2. Then output the label on a separate line starting with: Label: EXP or
Label: IMP or Label: NOTABU.

Follow this example format:
Reasoning: <reasoning>
Label: <EXP or IMP or NOTABU>

Here are a few examples of labeled texts:
Text: [Example explicit abusive text 1]
Reasoning: [Reasoning for EXP, including targeted/untargeted]
Label: EXP

Text: [Example implicit abusive text 1]
Reasoning: [Reasoning for IMP, including targeted/untargeted]
Label: IMP

Text: [Example non-abusive text 1]
Reasoning: [Reasoning for NOTABU, including targeted/untargeted]
Label: NOTABU
...

Now classify the following texts. ALWAYS output the instruction steps for EACH
of the N texts.

Texts:
1. [Text to classify 1]
2. [Text to classify 2]
...
N. [Text to classify N]
```

Strategy 5: Ternary - IHC

```
Classify the following texts into one of the following categories:
- explicit abuse (EXP)
- implicit abuse (IMP)
- not abusive (NOTABU)

Instructions for each text:
1. Write 1{2 short sentences explaining why the text is classified as EXP or
IMP or NOTABU. Consider if the text includes:
- white grievance
- incitement to violence
- inferiority language
- irony
- stereotypes and misinformation
- threatening and intimidation
2. Then output the label on a separate line starting with: Label: EXP or
Label: IMP or Label: NOTABU.

Follow this example format:
Reasoning: <reasoning>
Label: <EXP or IMP or NOTABU>

Here are a few examples of labeled texts:
Text: [Example explicit abusive text 1]
Reasoning: [Reasoning referencing IHC cues for EXP]
Label: EXP

Text: [Example implicit abusive text 1]
Reasoning: [Reasoning referencing IHC cues for IMP]
Label: IMP

Text: [Example non-abusive text 1]
Reasoning: [Reasoning referencing IHC cues for NOTABU]
Label: NOTABU
...

Now classify the following texts. ALWAYS output the instruction steps for EACH
of the N texts.

Texts:
1. [Text to classify 1]
2. [Text to classify 2]
...
N. [Text to classify N]
```

#	Label	Type	A	B	Topic	Tweet
Abusive (ABU)						
1	ABU	EXP	TIN	IND	Politics	@USER you are a lying corrupt traitor!!! Nobody wants to hear anymore of your lies!!! #DeepStateCorruption URL
2	ABU	IMP	UNT	–	Intelligence	@USER Gun Control means using both hands. And in this case means acquiring the correct target. You can't fix stupid.
3	ABU	EXP	TIN	GRP	Fat shaming	@USER "the sex symbol of rap" LMAOOO BITCH YOUre FAT aS FUCK ANYWAYS you tryna fight?
4	ABU	IMP	TIN	OTH	Politics	10kids from different women from this American jobless men? Nigerian men you are the bomb!!!
5	ABU	EXP	TIN	OTH	Politics	MAGA are PERVERTS AND DEMONS-PERIOD! NOTHING IS SACRED OR INNOCENT WITH THESE MUTTS OF SATAN URL
6	ABU	IMP	TIN	GRP	Politics	4 out of 10 British people are basically full-on racists. 4 out of 10 voters vote for the Conservatives. Coincidence!???!???
7	ABU	EXP	UNT	–	Activity	@USER What the fuck game are you watching?
8	ABU	IMP	TIN	IND	Politics	@USER Thank you for your fortitude and perseverance. Please give McConnell a kick in the butt from some of us conservatives.
9	ABU	EXP	TIN	IND	Gender	@USER If you go by anything other than he or she you are fucked in the head.
10	ABU	IMP	UNT	–	Gender	2 hours at midnight and she said she is not a call girl? -Awkward!
11	ABU	EXP	TIN	GRP	Politics	@USER @USER The FASCISTS are Anti-American Communist Scum like yourself!! Now go Fuck Yourself you piece of human garbage!! USA #MAGA ANTIFA are Terrorists
12	ABU	IMP	TIN	OTH	Racism	@USER Do you think she is afraid of pigs that are not the same colour as her?
Not Abusive (NOTABU)						
1	NOTABU	–	–	–	Politics	@USER What if knowing that abortion is at stake in November mobilizes pro life forces? Kavanaugh being denied makes it critical for conservatives to maintain the Senate. Because most (R) seats are not up for election it could be a blessing.
2	NOTABU	–	–	–	Human rights	(You are! My prayer is that you are blessed with someone you know like that also. Someone who feels your basic human rights are worth protecting. Especially your God given right to personal privacy.)
3	NOTABU	–	–	–	Education	@USER You should reach out to Paul.berger@USER He is teaching & researching these classes. & he is fierce & great! V. Ethical.
4	NOTABU	–	–	–	Censorship	@USER You guys have featured splendidrainco's etsy page and she is being censored by etsy now. You can't find her site through the search engine you have to have a direct link. And according to her instagram post they have been basically leaving her out to dry.
5	NOTABU	–	–	–	Compliment	@USER Aww Bradley you are so cute
6	NOTABU	–	–	–	Health	@USER So happy for you and your dad!!! My dad had the same kind of cancer but stage 4. He was diagnosed in 2016 and passed away this passed January. My mom is kicking leukemia's ass since Nov. of 2015 and she is doing awesome!
7	NOTABU	–	–	–	Sports	@USER @USER It's OK that you are confused. Paul Guenther is a proven DC and play caller. We are rebuilding and trying to remain competitive at the same time. We will have steady improvement
8	NOTABU	–	–	–	Entertainment	@USER I loved the House Bunny. And the remembering names voice was funny. My family and I watch Mom it's great. I'm a recovered addict or moderately recovered if you are one who counts pot lol. It's a very good show.
9	NOTABU	–	–	–	Politics	@USER @USER But Google can give Beijing every Americans phone number. You're kidding me right? MAGA 2020
10	NOTABU	–	–	–	Health	@USER sorry you are feeling sick hope you get better soon.
11	NOTABU	–	–	–	Politics	@USER That's not Southern California. It is considered Central California. Great how our gun control laws work here isn't it?
12	NOTABU	–	–	–	Entertainment	@USER It is so funny that she is scared of pigs. This could be a plot for your next children's book.

Table 1: Prompt examples for binary set-up labeled by label, subtasks, and topic.

#	Reasoning CoT	Reasoning CoT -Targeted	Reasoning CoT - IHC labels
Abusive (ABU)			
1	This message attacks a person directly using insults and strong anger.	This message targets an individual directly using insults and strong anger.	This message attacks a person directly using insults and strong anger. It includes threats and intimidation.
2	This message suggests violence and ends with an insult.	This message suggests violence and ends with an insult, but it is not directly targeted.	This message suggests violence and ends with an insult. It includes incitement to violence and irony.
3	This message includes profanity, body shaming, and a threat. It is rude and aggressive.	This message includes body shaming that is targeted toward a group.	This message includes profanity, body shaming, and a threat. It includes inferiority language and intimidation.
4	This message mocks Nigerian men using stereotypes and sarcasm. It is insulting and racist.	This message targets Nigerian men using stereotypes and sarcasm.	This message mocks Nigerian men using stereotypes and sarcasm. It includes stereotypes and misinformation.
5	This message uses extreme insults to attack a political movement. It is hateful and hostile.	This message uses extreme insults to target a political movement.	This message uses extreme insults to attack a political movement. It includes threats and dehumanizing language.
6	This message links a group to racism with a harmful stereotype and ends with a suggestive question. It is offensive.	This message targets a group by linking it to racism through a harmful stereotype and a suggestive question.	This message links a group to racism with a harmful stereotype and ends with a suggestive question. It includes stereotypes and misinformation.
7	This message asks a question in a disrespectful way and includes profanity.	This message asks a question in a disrespectful way and includes profanity, but remains untargeted.	This message asks a question in a disrespectful way and includes profanity. It includes irony and mild intimidation.
8	This message suggests physical aggression against someone. Even if partly joking, it is still hostile.	This message suggests physical aggression targeted at an individual.	This message suggests physical aggression against someone. Even if partly joking, it includes incitement to violence.
9	This message attacks people for their gender identity using profanity. It's demeaning.	This message targets an individual by attacking their gender identity using profanity.	This message attacks people for their gender identity using profanity. It includes inferiority language and intimidation.
10	This message implies a degrading assumption through a mocking question. It's suggestive and disrespectful.	This message implies a degrading assumption through a mocking question, but remains untargeted.	This message implies a degrading assumption through a mocking question. It includes irony and inferiority language.
11	This message is extremely aggressive and dehumanizing. It includes personal and political attacks.	This message uses personal and political insults to target a group.	This message is extremely aggressive and dehumanizing. It includes threats, intimidation, and hate toward a political movement.
12	This message implies a racial insult in a sarcastic tone and compares someone to animals.	This message implies a racial insult in a sarcastic way, targeting people based on skin color.	This message implies a racial insult in a sarcastic tone and compares someone to animals. It includes stereotypes and inferiority language.
Not Abusive (NOTABU)			
1	This message shares a political opinion without insults or hate.	This message shares a political opinion without any target.	This message shares a political opinion without insults or hate. It does not include any harmful framing.
2	This message includes wishes and prayers. It is kind and respectful toward someone.	This message includes wishes and prayers. It is kind and respectful toward someone, but not targeted.	This message includes wishes and prayers. It is kind and respectful toward someone. It does not include threats, stereotypes, or other harmful speech.
3	This message gives a helpful suggestion and ends by praising the suggested teacher.	This message gives a helpful suggestion and ends by praising the suggested teacher, but remains untargeted.	This message gives a helpful suggestion and ends by praising the suggested teacher. It does not include any harmful content.
4	This message expresses light frustration about how someone is treated but remains respectful.	This message expresses light frustration about how someone is treated but remains untargeted.	This message expresses light frustration about how someone is treated but remains respectful. It does not include stereotypes or insults.
5	This message is a simple, kind compliment.	This message is a simple, kind compliment and stays untargeted.	This message is a simple, kind compliment. It contains no negativity or bias.
6	This message shows support and tells a personal story about the health of family members.	This message shows support and tells a personal story about the health of family members. It does not include a target.	This message shows support and tells a personal story about the health of family members. It is empathetic and non-hostile.
7	This message shares a sports opinion in an affirming and supportive way.	This message shares a sports opinion in an affirming and supportive way without a target.	This message shares a sports opinion in an affirming and supportive way. It does not target or insult anyone.
8	This message talks about TV shows in a light and personal way.	This message talks about TV shows in a light and personal way without a target.	This message talks about TV shows in a light and personal way. It does not include stereotypes or aggressive tone.
9	This message expresses political skepticism while staying polite.	This message expresses political skepticism while staying polite and untargeted.	This message expresses political skepticism while staying polite. It does not include threats, hate, or misinformation.
10	This message shows support and wishes someone good health.	This message shows support and wishes someone good health, and stays untargeted.	This message shows support and wishes someone good health. It is friendly and caring.
11	This message is a sarcastic political opinion. It expresses light criticism but remains respectful.	This message is a sarcastic political opinion. It expresses light criticism but remains untargeted.	This message is a sarcastic political opinion. It expresses light criticism but remains respectful. It includes irony, but no harm.
12	This message discusses something funny and playfully suggests writing a book about it.	This message discusses something funny and playfully suggests writing a book about it, with no target.	This message discusses something funny and playfully suggests writing a book about it. It is not offensive.

Table 2: Reasoning annotations for Base, Targeted, and IHC in binary set-up.

#	Label	A	B	Topic	Tweet
Explicit Abuse (EXP)					
1	EXP	TIN	IND	Politics	@USER you are a lying corrupt traitor!!! Nobody wants to hear anymore of your lies!!! #DeepStateCorruption URL
2	EXP	TIN	GRP	Fat shaming	@USER ‘the sex symbol of rap’ LMAOOO BITCH YOUre FAT aS FUCK ANYWAYS you tryna fight?
3	EXP	TIN	OTH	Politics	MAGA are PERVERTS AND DEMONS—PERIOD! NOTHING IS SACRED OR INNOCENT WITH THESE MUTTS OF SATAN URL
4	EXP	UNT	–	Activity	@USER What the fuck game are you watching?
5	EXP	TIN	IND	Gender	@USER If you go by anything other than he or she you are fucked in the head.
6	EXP	TIN	GRP	Politics	@USER @USER The FASCISTS are Anti-American Communist Scum like yourself!! Now go fuck yourself you piece of human garbage!! USA #MAGA ANTIFA are Terrorists
7	EXP	TIN	OTH	Mental health	@USER @USER Are you bipolar? You keep saying different shit to people commenting on what u said
8	EXP	UNT	–	Insult	@USER So what ur saying is that it is in fact chicken shit and you are eating chicken shit URL
Implicit Abuse (IMP)					
1	IMP	UNT	–	Intelligence	@USER Gun Control means using both hands. And in this case means acquiring the correct target. You can’t fix stupid.
2	IMP	TIN	OTH	Racism	10kids from different women from this American jobless men? Nigerian men you are the bomb!!!
3	IMP	TIN	GRP	Politics	4 out of 10 British people are basically full-on racists. 4 out of 10 voters vote for the Conservatives. Coincidence!???!???
4	IMP	TIN	IND	Politics	@USER Thank you for your fortitude and perseverance. Please give McConnell a kick in the butt from some of us conservatives.
5	IMP	UNT	–	Gender	2 hours at midnight and she said she is not a call girl? -Awkward!
6	IMP	TIN	OTH	Racism	@USER Do you think she is afraid of pigs that are not the same colour as her?
7	IMP	TIN	GRP	Fat shaming	@USER Love how fat they all are while chowing down on cake and cookies. That’s the future liberals want.
8	IMP	TIN	IND	Politics	@USER Absolutely should be a nail in the coffin.... for Hillary
Not Abusive (NOTABU)					
1	NOTABU	–	–	Politics	@USER What if knowing that abortion is at stake in November mobilizes pro life forces? Kavanaugh being denied makes it critical for conservatives to maintain the Senate. Because most (R) seats are not up for election it could be a blessing.
2	NOTABU	–	–	Human rights	(You are! My prayer is that you are blessed with someone you know like that also. Someone who feels your basic human rights are worth protecting. Especially your God given right to personal privacy.)
3	NOTABU	–	–	Education	@USER You should reach out to Paul.berger@USER He is teaching & researching these classes. & he is fierce & great! V. Ethical.
4	NOTABU	–	–	Censorship	@USER You guys have featured splendidrainco’s etsy page and she is being censored by etsy now. You can’t find her site through the search engine you have to have a direct link. And according to her instagram post they have been basically leaving her out to dry.
5	NOTABU	–	–	Compliment	@USER Aww Bradley you are so cute
6	NOTABU	–	–	Health	@USER So happy for you and your dad!!! My dad had the same kind of cancer but stage 4. He was diagnosed in 2016 and passed away this passed January. My mom is kicking leukemia’s ass since Nov. of 2015 and she is doing awesome!
7	NOTABU	–	–	Sports	@USER @USER It’s OK that you are confused. Paul Guenther is a proven DC and play caller. We are rebuilding and trying to remain competitive at the same time. We will have steady improvement
8	NOTABU	–	–	Entertainment	@USER I loved the House Bunny. And the remembering names voice was funny. My family and I watch Mom it’s great. I’m a recovered addict or moderately recovered if you are one who counts pot lol. It’s a very good show.

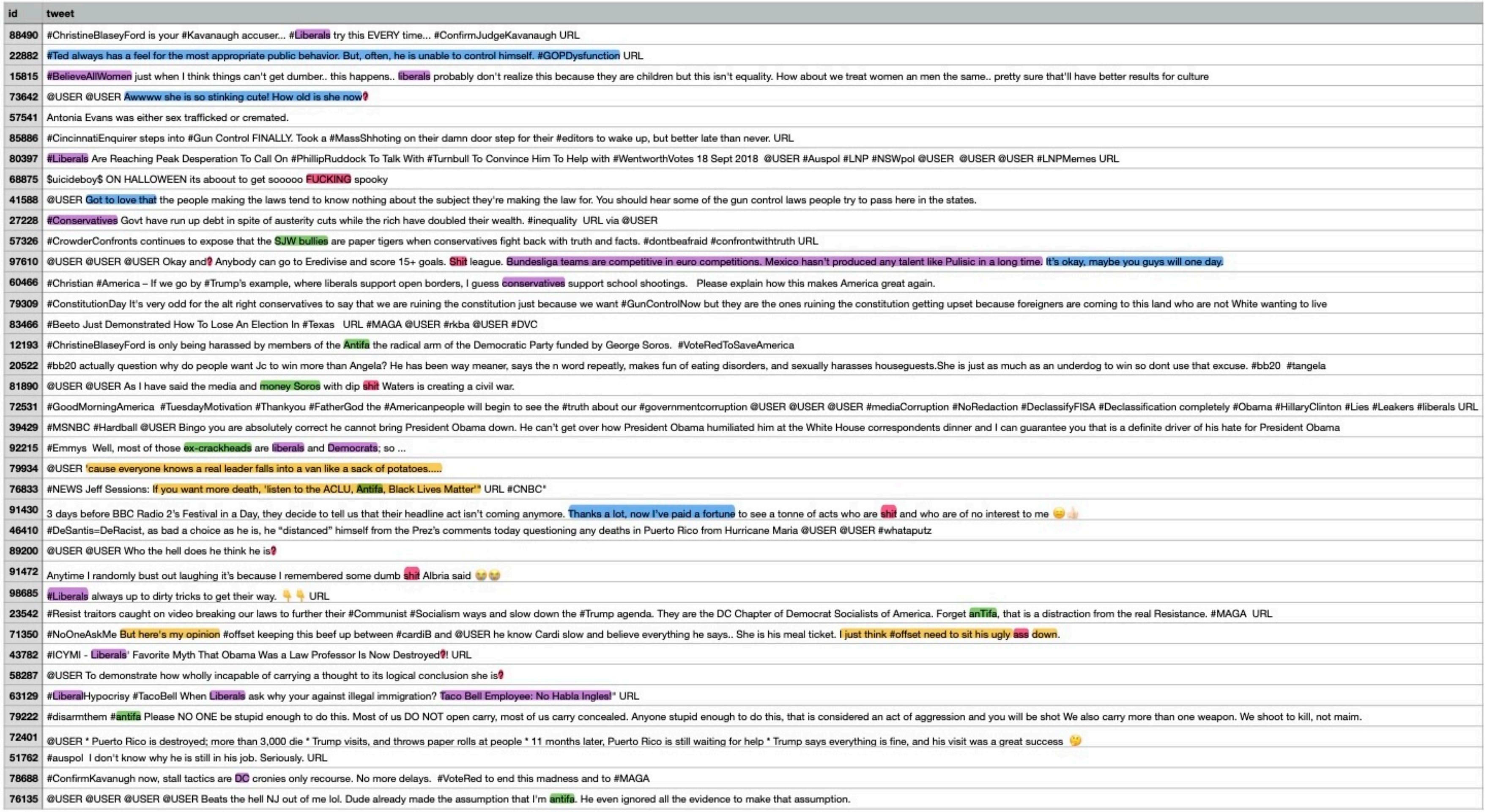
Table 3: Prompt examples for ternary set-up labeled by label, subtasks, and topic.

#	Reasoning CoT	Reasoning CoT -Targeted	Reasoning CoT - IHC labels
Explicit Abuse (EXP)			
1	This message attacks a person directly using insults and strong anger.	This message targets an individual directly using insults and strong anger.	This message attacks a person directly using insults and strong anger. It includes threats and intimidation.
2	This message includes profanity, body shaming, and a threat. It is rude and aggressive.	This message includes body shaming that is targeted toward a group.	This message includes profanity, body shaming, and a threat. It includes inferiority language and intimidation.
3	This message uses extreme insults to attack a political movement. It is hateful and hostile.	This message uses extreme insults to target a political movement.	This message uses extreme insults to attack a political movement. It includes threats and dehumanizing language.
4	This message asks a question in a disrespectful way and includes profanity.	This message asks a question in a disrespectful way and includes profanity, but remains untargeted.	This message asks a question in a disrespectful way and includes profanity. It includes irony and mild intimidation.
5	This message attacks people for their gender identity using profanity. It's demeaning.	This message targets an individual by attacking their gender identity using profanity.	This message attacks people for their gender identity using profanity. It includes inferiority language and intimidation.
6	This message is extremely aggressive and dehumanizing. It includes personal and political attacks.	This message uses personal and political insults to target a group.	This message is extremely aggressive and dehumanizing. It includes threats, intimidation, and hate toward a political movement.
7	This message insults someone by implying a mental health condition in a mocking way. It uses profanity and is demeaning.	This message targets people with a mental illness in a mocking and disrespectful way.	This message uses profanity and mocks a mental health condition to insult someone. It includes inferiority language and is directly demeaning.
8	This message uses crude language to describe what someone is saying. It's vulgar, insulting, and disrespectful.	This message uses vulgar language to insult someone, but it remains untargeted.	This message uses vulgar and dehumanizing language to insult someone in a crude way. It shows strong disrespect but is not directly threatening.
Implicit Abuse (IMP)			
1	This message suggests violence and ends with an insult.	This message suggests violence and ends with an insult, but it is not directly targeted.	This message suggests violence and ends with an insult. It includes incitement to violence and irony.
2	This message mocks Nigerian men using stereotypes and sarcasm. It is insulting and racist.	This message targets Nigerian men using stereotypes and sarcasm.	This message mocks Nigerian men using stereotypes and sarcasm. It includes stereotypes and misinformation.
3	This message links a group to racism with a harmful stereotype and ends with a suggestive question. It is offensive.	This message targets a group by linking it to racism through a harmful stereotype and a suggestive question.	This message links a group to racism with a harmful stereotype and ends with a suggestive question. It includes stereotypes and misinformation.
4	This message suggests physical aggression against someone. Even if partly joking, it is still hostile.	This message suggests physical aggression targeted at an individual.	This message suggests physical aggression against someone. Even if partly joking, it includes incitement to violence.
5	This message implies a degrading assumption through a mocking question. It's suggestive and disrespectful.	This message implies a degrading assumption through a mocking question, but remains untargeted.	This message implies a degrading assumption through a mocking question. It includes irony and inferiority language.
6	This message implies a racial insult in a sarcastic tone and compares someone to animals.	This message implies a racial insult in a sarcastic way, targeting people based on skin color.	This message implies a racial insult in a sarcastic tone and compares someone to animals. It includes stereotypes and inferiority language.
7	This message insults a political group through body shaming and sarcasm. It's demeaning and indirectly hostile.	This message targets a political group through body shaming and sarcasm.	This message uses body shaming and sarcasm to insult a political group. It includes irony and inferiority language.
8	This message uses a metaphor associated with death to refer to a political figure. It is ominous and implicitly threatening.	This message suggests harm toward an individual through a metaphor associated with death.	This message uses a metaphor associated with death to suggest harm toward political figure. It implies incitement to violence and includes threats and intimidation.
Not Abusive (NOTABU)			
1	This message shares a political opinion without insults or hate.	This message shares a political opinion without any target.	This message shares a political opinion without insults or hate. It does not include any harmful framing.
2	This message includes wishes and prayers. It is kind and respectful toward someone.	This message includes wishes and prayers. It is kind and respectful toward someone, but not targeted.	This message includes wishes and prayers. It is kind and respectful toward someone. It does not include threats, stereotypes, or other harmful speech.
3	This message gives a helpful suggestion and ends by praising the suggested teacher.	This message gives a helpful suggestion and ends by praising the suggested teacher, but remains untargeted.	This message gives a helpful suggestion and ends by praising the suggested teacher. It does not include any harmful content.
4	This message expresses light frustration about how someone is treated but remains respectful.	This message expresses light frustration about how someone is treated but remains untargeted.	This message expresses light frustration about how someone is treated but remains respectful. It does not include stereotypes or insults.
5	This message is a simple, kind compliment.	This message is a simple, kind compliment and stays untargeted.	This message is a simple, kind compliment. It contains no negativity or bias.
6	This message shows support and tells a personal story about the health of family members.	This message shows support and tells a personal story about the health of family members. It does not include a target.	This message shows support and tells a personal story about the health of family members. It is empathetic and non-hostile.
7	This message shares a sports opinion in an affirming and supportive way.	This message shares a sports opinion in an affirming and supportive way without a target.	This message shares a sports opinion in an affirming and supportive way. It does not target or insult anyone.
8	This message talks about TV shows in a light and personal way.	This message talks about TV shows in a light and personal way without a target.	This message talks about TV shows in a light and personal way. It does not include stereotypes or aggressive tone.

Table 4: Reasoning annotations for Base, Targeted, and IHC in ternary set-up.

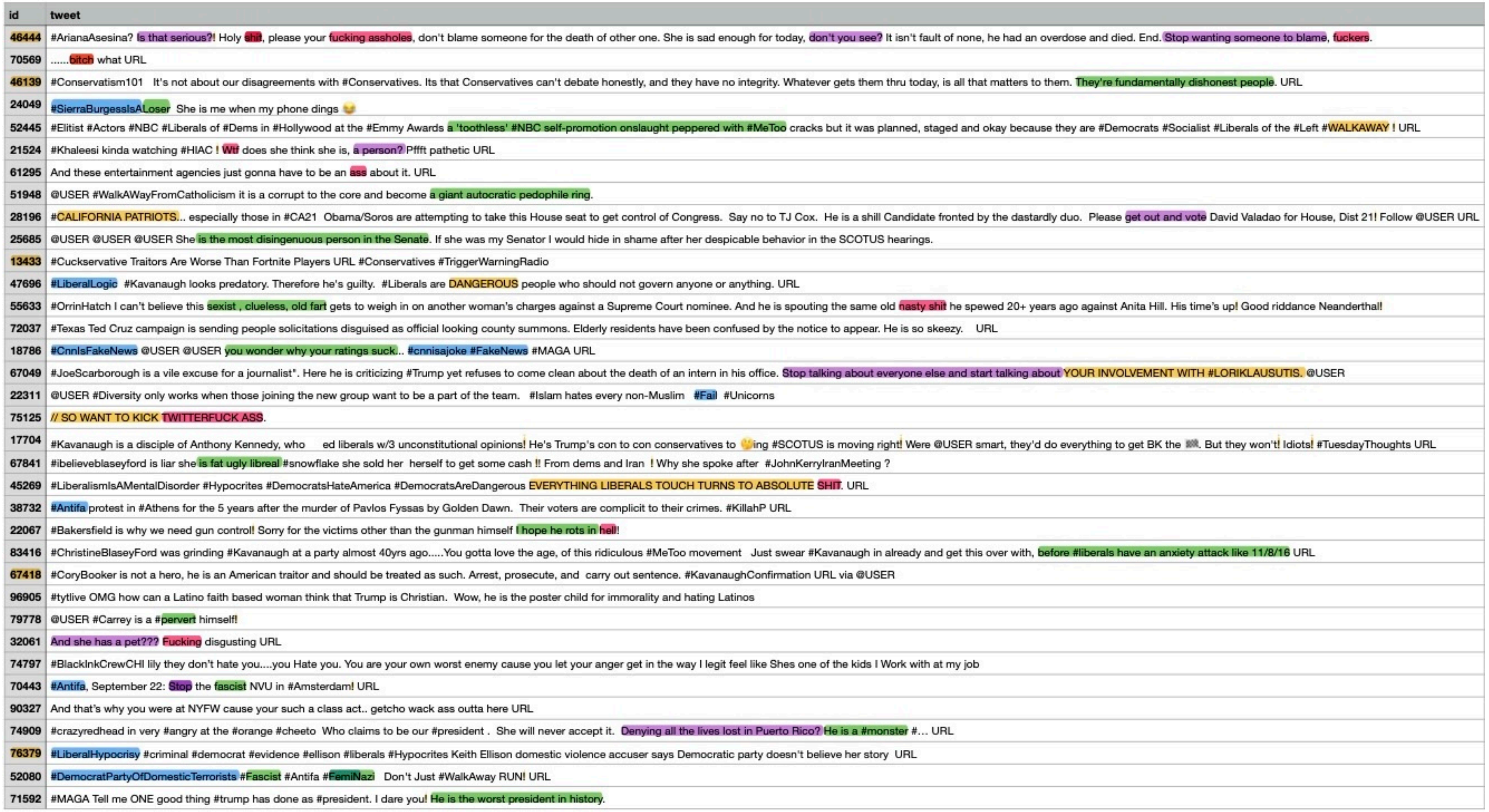
Appendix C

This appendix shows the marks of examples for the manual part of the error analysis. Figure 1 shows the ones for IMP as NOTABU. Figure 2 shows the ones for EXP as IMP.



id	tweet
88490	#ChristineBlaseyFord is your #Kavanaugh accuser... #Liberals try this EVERY time... #ConfirmJudgeKavanaugh URL
22882	#Ted always has a feel for the most appropriate public behavior. But, often, he is unable to control himself. #GOPDysfunction URL
15815	#BelieveAllWomen just when I think things can't get dumber.. this happens.. liberals probably don't realize this because they are children but this isn't equality. How about we treat women an men the same.. pretty sure that'll have better results for culture
73642	@USER @USER Awwww she is so stinking cute! How old is she now?
57541	Antonia Evans was either sex trafficked or cremated.
85886	#CincinnatiEnquirer steps into #Gun Control FINALLY. Took a #MassShhoting on their damn door step for their #editors to wake up, but better late than never. URL
80397	#Liberals Are Reaching Peak Desperation To Call On #PhillipRuddock To Talk With #Turnbull To Convince Him To Help with #WentworthVotes 18 Sept 2018 @USER #Auspol #LNP #NSWpol @USER @USER @USER #LNPMemes URL
68875	$uicideboy$ ON HALLOWEEN its aboout to get sooooo FUCKING spooky
41588	@USER Got to love that the people making the laws tend to know nothing about the subject they're making the law for. You should hear some of the gun control laws people try to pass here in the states.
27228	#Conservatives Govt have run up debt in spite of austerity cuts while the rich have doubled their wealth. #inequality URL via @USER
57326	#CrowderConfronts continues to expose that the SJW bullies are paper tigers when conservatives fight back with truth and facts. #dontbeafraid #confrontwithtruth URL
97610	@USER @USER @USER Okay and? Anybody can go to Eredivise and score 15+ goals. Shit league. Bundesliga teams are competitive in euro competitions. Mexico hasn't produced any talent like Pulisic in a long time. It's okay, maybe you guys will one day.
60466	#Christian #America – If we go by #Trump's example, where liberals support open borders, I guess conservatives support school shootings. Please explain how this makes America great again.
79309	#ConstitutionDay It's very odd for the alt right conservatives to say that we are ruining the constitution just because we want #GunControlNow but they are the ones ruining the constitution getting upset because foreigners are coming to this land who are not White wanting to live
83466	#Beeto Just Demonstrated How To Lose An Election In #Texas URL #MAGA @USER #rkba @USER #DVC
12193	#ChristineBlaseyFord is only being harassed by members of the Antifa the radical arm of the Democratic Party funded by George Soros. #VoteRedToSaveAmerica
20522	#bb20 actually question why do people want Jc to win more than Angela? He has been way meaner, says the n word repeatly, makes fun of eating disorders, and sexually harasses houseguests.She is just as much as an underdog to win so dont use that excuse. #bb20 #tangela
81890	@USER @USER As I have said the media and money Soros with dip shit Waters is creating a civil war.
72531	#GoodMorningAmerica #TuesdayMotivation #Thankyou #FatherGod the #Americanpeople will begin to see the #truth about our #governmentcorruption @USER @USER @USER #mediaCorruption #NoRedaction #DeclassifyFISA #Declassification completely #Obama #HillaryClinton #Lies #Leakers #liberals URL
39429	#MSNBC #Hardball @USER Bingo you are absolutely correct he cannot bring President Obama down. He can't get over how President Obama humiliated him at the White House correspondents dinner and I can guarantee you that is a definite driver of his hate for President Obama
92215	#Emmys Well, most of those ex-crackheads are liberals and Democrats; so ...
79934	@USER 'cause everyone knows a real leader falls into a van like a sack of potatoes.....
76833	#NEWS Jeff Sessions: If you want more death, 'listen to the ACLU, Antifa, Black Lives Matter'" URL #CNBC"
91430	3 days before BBC Radio 2's Festival in a Day, they decide to tell us that their headline act isn't coming anymore. Thanks a lot, now I've paid a fortune to see a tonne of acts who are shit and who are of no interest to me 😑👍🏻
46410	#DeSantis=DeRacist, as bad a choice as he is, he "distanced" himself from the Prez's comments today questioning any deaths in Puerto Rico from Hurricane Maria @USER @USER #whataputz
89200	@USER @USER Who the hell does he think he is?
91472	Anytime I randomly bust out laughing it's because I remembered some dumb shit Albria said 😭😭
98685	#Liberals always up to dirty tricks to get their way. 👇👇 URL
23542	#Resist traitors caught on video breaking our laws to further their #Communist #Socialism ways and slow down the #Trump agenda. They are the DC Chapter of Democrat Socialists of America. Forget anTifa, that is a distraction from the real Resistance. #MAGA URL
71350	#NoOneAskMe But here's my opinion #offset keeping this beef up between #cardiB and @USER he know Cardi slow and believe everything he says.. She is his meal ticket. I just think #offset need to sit his ugly ass down.
43782	#ICYMI - Liberals' Favorite Myth That Obama Was a Law Professor Is Now Destroyed?! URL
58287	@USER To demonstrate how wholly incapable of carrying a thought to its logical conclusion she is?
63129	#LiberalHypocrisy #TacoBell When Liberals ask why your against illegal immigration? Taco Bell Employee: No Habla Ingles!" URL
79222	#disarmthem #antifa Please NO ONE be stupid enough to do this. Most of us DO NOT open carry, most of us carry concealed. Anyone stupid enough to do this, that is considered an act of aggression and you will be shot We also carry more than one weapon. We shoot to kill, not maim.
72401	@USER * Puerto Rico is destroyed; more than 3,000 die * Trump visits, and throws paper rolls at people * 11 months later, Puerto Rico is still waiting for help * Trump says everything is fine, and his visit was a great success 🤔
51762	#auspol I don't know why he is still in his job. Seriously. URL
78688	#ConfirmKavanugh now, stall tactics are DC cronies only recourse. No more delays. #VoteRed to end this madness and to #MAGA
76135	@USER @USER @USER @USER Beats the hell NJ out of me lol. Dude already made the assumption that I'm antifa. He even ignored all the evidence to make that assumption.

Figure 1: Markings for error analysis: IMP as NOTABU. Pink = Profanity and rhetorical questions, Purple = Stereotyping and generalizations, Pink '?' = Rhetorical questions. Blue = Sarcasm and irony / false politeness and passive-aggressiveness, Green = Coded or euphemistic language, Yellow = Abuse masked as moral or opinion.



id	tweet
46444	#ArianaAsesina? Is that serious?! Holy shit, please your fucking assholes, don't blame someone for the death of other one. She is sad enough for today, don't you see? It isn't fault of none, he had an overdose and died. End. Stop wanting someone to blame, fuckers.
70569	......bitch what URL
46139	#Conservatism101 It's not about our disagreements with #Conservatives. Its that Conservatives can't debate honestly, and they have no integrity. Whatever gets them thru today, is all that matters to them. They're fundamentally dishonest people. URL
24049	#SierraBurgessIsALoser She is me when my phone dings 😂
52445	#Elitist #Actors #NBC #Liberals of #Dems in #Hollywood at the #Emmy Awards a 'toothless' #NBC self-promotion onslaught peppered with #MeToo cracks but it was planned, staged and okay because they are #Democrats #Socialist #Liberals of the #Left #WALKAWAY ! URL
21524	#Khaleesi kinda watching #HIAC ! Wtf does she think she is, a person? Pffft pathetic URL
61295	And these entertainment agencies just gonna have to be an ass about it. URL
51948	@USER #WalkAWayFromCatholicism it is a corrupt to the core and become a giant autocratic pedophile ring.
28196	#CALIFORNIA PATRIOTS... especially those in #CA21 Obama/Soros are attempting to take this House seat to get control of Congress. Say no to TJ Cox. He is a shill Candidate fronted by the dastardly duo. Please get out and vote David Valadao for House, Dist 21! Follow @USER URL
25685	@USER @USER @USER She is the most disingenuous person in the Senate. If she was my Senator I would hide in shame after her despicable behavior in the SCOTUS hearings.
13433	#Cuckservative Traitors Are Worse Than Fortnite Players URL #Conservatives #TriggerWarningRadio
47696	#LiberalLogic #Kavanaugh looks predatory. Therefore he's guilty. #Liberals are DANGEROUS people who should not govern anyone or anything. URL
55633	#OrrinHatch I can't believe this sexist , clueless, old fart gets to weigh in on another woman's charges against a Supreme Court nominee. And he is spouting the same old nasty shit he spewed 20+ years ago against Anita Hill. His time's up! Good riddance Neanderthal!
72037	#Texas Ted Cruz campaign is sending people solicitations disguised as official looking county summons. Elderly residents have been confused by the notice to appear. He is so skeezy. URL
18786	#CnnIsFakeNews @USER @USER you wonder why your ratings suck... #cnnisajoke #FakeNews #MAGA URL
67049	#JoeScarborough is a vile excuse for a journalist". Here he is criticizing #Trump yet refuses to come clean about the death of an intern in his office. Stop talking about everyone else and start talking about YOUR INVOLVEMENT WITH #LORIKLAUSUTIS. @USER
22311	@USER #Diversity only works when those joining the new group want to be a part of the team. #Islam hates every non-Muslim #Fail #Unicorns
75125	// SO WANT TO KICK TWITTERFUCK ASS.
17704	#Kavanaugh is a disciple of Anthony Kennedy, who ed liberals w/3 unconstitutional opinions! He's Trump's con to con conservatives to 😠ing #SCOTUS is moving right! Were @USER smart, they'd do everything to get BK the 🏁. But they won't! Idiots! #TuesdayThoughts URL
67841	#ibelieveblaseyford is liar she is fat ugly libreal #snowflake she sold her herself to get some cash !! From dems and Iran ! Why she spoke after #JohnKerryIranMeeting ?
45269	#LiberalismIsAMentalDisorder #Hypocrites #DemocratsHateAmerica #DemocratsAreDangerous EVERYTHING LIBERALS TOUCH TURNS TO ABSOLUTE SHIT. URL
38732	#Antifa protest in #Athens for the 5 years after the murder of Pavlos Fyssas by Golden Dawn. Their voters are complicit to their crimes. #KillahP URL
22067	#Bakersfield is why we need gun control! Sorry for the victims other than the gunman himself I hope he rots in hell!
83416	#ChristineBlaseyFord was grinding #Kavanaugh at a party almost 40yrs ago.....You gotta love the age, of this ridiculous #MeToo movement Just swear #Kavanaugh in already and get this over with, before #liberals have an anxiety attack like 11/8/16 URL
67418	#CoryBooker is not a hero, he is an American traitor and should be treated as such. Arrest, prosecute, and carry out sentence. #KavanaughConfirmation URL via @USER
96905	#tytlive OMG how can a Latino faith based woman think that Trump is Christian. Wow, he is the poster child for immorality and hating Latinos
79778	@USER #Carrey is a #pervert himself!
32061	And she has a pet??? Fucking disgusting URL
74797	#BlackInkCrewCHI lily they don't hate you....you Hate you. You are your own worst enemy cause you let your anger get in the way I legit feel like Shes one of the kids I Work with at my job
70443	#Antifa, September 22: Stop the fascist NVU in #Amsterdam! URL
90327	And that's why you were at NYFW cause your such a class act.. getcho wack ass outta here URL
74909	#crazyredhead in very #angry at the #orange #cheeto Who claims to be our #president . She will never accept it. Denying all the lives lost in Puerto Rico? He is a #monster #... URL
76379	#LiberalHypocrisy #criminal #democrat #evidence #ellison #liberals #Hypocrites Keith Ellison domestic violence accuser says Democratic party doesn't believe her story URL
52080	#DemocratPartyOfDomesticTerrorists #Fascist #Antifa #FemiNazi Don't Just #WalkAway RUN! URL
71592	#MAGA Tell me ONE good thing #trump has done as #president. I dare you! He is the worst president in history.

Figure 2: Markings for error analysis: EXP as IMP. Pink = Profanity, Yellow = Uppercase words, exclamations, and moralistic abuse, Yellow id number = Abuse makes as moral or opinion, Purple = Rhetorical Questions, Green = Direct insults and slurs / direct references, Blue = Embedded hashtags.

References

P. Badjatiya, S. Gupta, M. Gupta, and V. Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, WWW '17 Companion, page 759–760. ACM Press, 2017. doi: 10.1145/3041021.3054223. URL http://dx.doi.org/10.1145/3041021.3054223.

V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, and S. M. Mohammad, editors, *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2007. URL https://aclanthology.org/S19-2007/.

T. Caselli, V. Basile, J. Mitrović, I. Kartoziya, and M. Granitzer. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.760/.

T. Caselli, V. Basile, J. Mitrović, and M. Granitzer. HateBERT: Retraining BERT for abusive language detection in English. In A. Mostafazadeh Davani, D. Kiela, M. Lambert, B. Vidgen, V. Prabhakaran, and Z. Waseem, editors, *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.woah-1.3. URL https://aclanthology.org/2021.woah-1.3/.

T. Davidson, D. Warmsley, M. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)*, pages 512–515, Montreal, Canada, 2017.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. Belding. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM)*, June 2018.

M. ElSherief, C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. De Choudhury, and D. Yang. Latent hatred: A benchmark for understanding implicit hate speech. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.29. URL https://aclanthology.org/2021.emnlp-main.29/.

A. G. et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

P. Fortuna and S. Nunes. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4), July 2018. ISSN 0360-0300. doi: 10.1145/3232676. URL https://doi.org/10.1145/3232676.

A.-M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, pages 512–515, Stanford, CA, USA, 2018. AAAI Press. doi: 10.1609/icwsm.v12i1.14991.

K. Guo, A. Hu, J. Mu, Z. Shi, Z. Zhao, N. Vishwamitra, and H. Hu. An investigation of large language models for real-world hate speech detection. In *Proceedings of the 2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1568–1573. IEEE, 2023. doi: 10.1109/ICMLA58977.2023.00237.

L. Han and H. Tang. Designing of prompts for hate speech recognition with in-context learning. In *2022 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 319–320, 2022. doi: 10.1109/CSCI58124.2022.00063.

T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.234. URL https://aclanthology.org/2022.acl-long.234/.

P. He, X. Liu, J. Gao, and W. Chen. Deberta: Decoding-enhanced bert with disentangled attention, 2021. URL https://arxiv.org/abs/2006.03654.

F. Huang, H. Kwak, and J. An. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23, page 294–297. ACM, Apr. 2023. doi: 10.1145/3543873.3587368. URL http://dx.doi.org/10.1145/3543873.3587368.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. URL https://nlp.stanford.edu/IR-book/.

B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875, 2021.

N. B. Ocampo, E. Sviridova, E. Cabrio, and S. Villata. An in-depth analysis of implicit and subtle hate speech messages. In A. Vlachos and I. Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.147. URL https://aclanthology.org/2023.eacl-main.147/.

OpenAI. Gpt-3.5 technical overview, 2023. URL https://platform.openai.com/docs/models/gpt-3-5.

OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos. Semeval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1526–1545, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.semeval-1.170.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. sklearn.metrics.classification_report, 2024. URL https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html. Accessed: 2025-05-21.

P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, and J. Pierrehumbert. HateCheck: Functional tests for hate speech detection models. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.4. URL https://aclanthology.org/2021.acl-long.4/.

S. Roy, A. Harshvardhan, A. Mukherjee, and P. Saha. Probing LLMs for hate speech detection: strengths and vulnerabilities. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6116–6128, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.407. URL https://aclanthology.org/2023.findings-emnlp.407/.

H. Saleh, A. Alhothali, and K. Moria. Detection of hate speech using bert and hate speech word embedding with deep model. *Applied Artificial Intelligence*, 37(1), 2023. doi: 10.1080/08839514.2023.2166719.

B. Vidgen, D. Nguyen, H. Margetts, P. Rossini, and R. Tromble. Introducing CAD: the contextual abuse dataset. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.182. URL https://aclanthology.org/2021.naacl-main.182/.

W. Warner and J. Hirschberg. Detecting hate speech on the world wide web. In S. O. Sood, M. Nagarajan, and M. Gamon, editors, *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada, June 2012. Association for Computational Linguistics. URL https://aclanthology.org/W12-2103/.

Z. Waseem and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In J. Andreas, E. Choi, and A. Lazaridou, editors, *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-2013. URL https://aclanthology.org/N16-2013/.

Z. Waseem, T. Davidson, D. Warmsley, and I. Weber. Understanding abuse: A typology of abusive language detection subtasks. In Z. Waseem, W. H. K. Chung, D. Hovy, and J. Tetreault, editors, *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada, Aug. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3012. URL https://aclanthology.org/W17-3012/.

A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. Predicting the type and target of offensive posts in social media. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1144. URL https://aclanthology.org/N19-1144/.

C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang. Can large language models transform computational social science? *Computational Linguistics*, 50(1): 237–291, Mar. 2024. doi: 10.1162/coli_a_00502. URL https://aclanthology.org/2024.cl-1.8/.