Interpretable Hate Speech Detection via Large Language Model-extracted Rationales

by

Ayushi Nirmal

A Thesis Presented in Partial Fulfillment
of the Requirement for the Degree
Master of Science

Approved April 2024 by the
Graduate Supervisory Committee:

Huan Liu, Chair
Hasan Davulcu
Hua Wei

ARIZONA STATE UNIVERSITY

May 2024

ABSTRACT

Social media platforms have become widely used for open communication, yet their lack of moderation has led to the proliferation of harmful content, including hate speech. Manual monitoring of such vast amounts of user-generated data is impractical, thus necessitating automated hate speech detection methods. Pre-trained language models have been proven to possess strong base capabilities, which not only excel at in-distribution language modeling but also show powerful abilities in out-of-distribution language modeling, transfer learning and few-shot learning. However, these models operate as complex function approximators, mapping input text to a hate speech classification, without providing any insights into the reasoning behind their predictions. Hence, existing methods often lack transparency, hindering their effectiveness, particularly in sensitive content moderation contexts. Recent efforts have been made to integrate their capabilities with large language models like ChatGPT and Llama2, which exhibit reasoning capabilities and broad knowledge utilization. This thesis explores leveraging the reasoning abilities of large language models to enhance the interpretability of hate speech detection. A novel framework is proposed that utilizes state-of-the-art Large Language Models (LLMs) to extract interpretable rationales from input text, highlighting key phrases or sentences relevant to hate speech classification. By incorporating these rationale features into a hate speech classifier, the framework inherently provides transparent and interpretable results. This approach combines the language understanding prowess of LLMs with the discriminative power of advanced hate speech classifiers, offering a promising solution to the challenge of interpreting automated hate speech detection models.

Keywords: Social Media, Hate Speech, Large Language Models, Rationale Extraction, Interpretability.

To my family, friends for always supporting me, and making me believe in myself.

ACKNOWLEDGEMENTS

I am grateful to my advisor, Dr. Huan Liu, for presenting me with the invaluable opportunity to delve into research in this field and for providing indispensable guidance throughout my master's journey. I extend my sincere gratitude to Bohan Jiang, Amrita Bhattacharjee, and Paras Sheth, who served as exceptional mentors, offering unwavering support at every step.

I would also like to express my heartfelt gratitude to all the members of the DMML lab group, especially Garima, Suraj, Hirthik, Nayoung, Teila, Zeyad, Faisal, Mansooreh, Tharindu, Ujun, Kaize, Zhen, Ali, Anique, Raha, Chengshuai, Saketh, and Saurabh who were always willing to lend a sympathetic ear and provide insightful perspectives to my numerous inquiries. I am thankful to my friends at ASU – Shubhodeep, Aseem, Pragya, Jeet, Darsh, and Ramya – for making this journey truly memorable. I owe a special debt of gratitude to my mother, Hemlata, my father, S.K. Nirmal, and my brother, Ayush Nirmal, for their constant encouragement, love, and unwavering support, which motivated me.

I sincerely thank Dr. Huan Liu for providing the funding support for my master's.

Finally, I sincerely appreciate and am thankful to Dr. Hasan Davulcu and Dr. Hua Wei for readily agreeing to serve on my thesis committee. This work would not have been achievable without the support and encouragement of everyone mentioned here. I extend my heartfelt gratitude to all.

TABLE OF CONTENTS

Page

LIST OF TABLES vi

LIST OF FIGURES vii

CHAPTER

1 INTRODUCTION 1

2 RELATED WORKS 4

2.1 Hate Speech Detection 4

2.1.1 Traditional Hate Speech Detection Approaches 5

2.1.2 Machine Learning Approaches 5

2.1.3 Ethical Considerations and Bias Mitigation 6

2.2 Large Language Models 10

2.2.1 Feature Extraction with LLMs 11

2.2.2 Significance of LLMs as Feature Extractors 12

2.2.3 LLMs Understand Semantics 13

2.2.4 LLMs Understand Contextual Information 13

3 FRAMEWORK 15

3.1 Motivation 15

3.2 Model Architecture 17

3.2.1 LLM as Rationales Extractor 17

3.2.2 Hate Speech Detector as Embedding Module 19

3.2.3 Feature Embedding Model 20

3.2.4 Embedding Fusion & Classification 20

3.3 Datasets 22

3.4 Baselines 24

3.5 Experiments 25

CHAPTER Page

4 RESULTS 26
4.1 Performance of ChatGPT on the hate speech detection task 27
4.2 Goodness of ChatGPT extracted features or rationales 29
4.3 Hate speech detector performance after training with extracted rationales 32
4.4 Interpretability with respect to human annotations 33
4.5 Modifying the hate speech detector and feature embedding models . 34
5 CONCLUSION AND FUTURE WORK 37
5.1 Summary 37
5.2 Future Work 38
5.3 Limitations 39
REFERENCES 40
APPENDIX
A CSS EXPLORATORY RESEARCH IN THE FIELD OF LLMs 47

LIST OF TABLES

Table Page

3.1 Dataset Statistics for Hate Speech Datasets from Different Social Media Platforms. 23

4.1 Examples of Input Text, Prompt, and ChatGPT's Response for a Data Sample from the Twitter Dataset. 28

4.2 Evaluation Results for Our Shield Framework Vs. The Baseline Models. Implicit HS Refers to the Implicit Hate Speech Corpus. Values in **Bold** Denote the Best Performance, and Underlined Values Denotes the Second-best Performance. 29

4.3 Examples of Input Text along with the LLM-extracted Features and Rationales. Rationales Are in Blue, Derogatory Language Is in Red, Cuss Words Are in Teal. 30

4.4 Similarity Between HateXplain Human Explanations and LLM-extracted Features/Rationales. 32

4.5 Interpretability Metrics: Similarity Between Human Annotated Rationales and Model Attention Tokens rationales for the HateXplain dataset 33

4.6 Analysis of HSD and FE Model Choices in the SHIELD Framework. HSD: Hate Speech Detector, FE: Feature Embedding Model. The Original SHIELD Framework Has Hatebert as the Hate Speech Detector and Bert-base-uncased as the Feature Embedding Model. Numbers in Bold Denote Best Performing Model Variant for Each Dataset. 36

LIST OF FIGURES

Figure Page

2.1 Hierarchy of Hate Speech Concepts (Alkomah and Ma, 2022) 7

3.1 An Overview of the Proposed Framework Architecture. 17

4.1 Examples with Both LLM-Annotated and Human-Annotated Rationales. Overlap Is in Purple. 31

A.1 Overview of Disinformation Generation and Detection Using Chatgpt. We Input Human-crafted Disinformation (Left) along with Distinct Prompts to Produce Three Separate Llm-generated Disinformation Datasets (Center). We Subsequently Evaluate the Efficacy of the Disinformation Detection System (Right) Against Llm-generated Disinformation (Jiang *et al.*, 2024). 49

A.2 The Migration Flow Between Twitter and Its Alternatives: Mastodon, Bluesky, and Threads. The Dashed Lines Represent the Shift of User Attention Across These Platforms (Jeong *et al.*, 2024). 50

Chapter 1

INTRODUCTION

Content Warning: This document contains content that some may find disturbing or offensive, including content that is discriminative, hateful, or violent in nature.

Social media platforms have emerged as a space for individuals from diverse cultural and geographical backgrounds to engage in content sharing and discussions. While these online conversations facilitate the exchange of information, they can sometimes escalate into unpleasant confrontations and bigoted arguments, leading to the proliferation of hate speech on these platforms. Hate speech refers to deliberate and purposeful public communication that expresses hatred, disdain, or contempt towards an individual or group based on their social attributes, such as gender or race (Nockleby, 1994; Perera *et al.*, 2023). In extreme cases, hate speech may often lead to real world harms such as hate crimes, for example the anti-Asian hate crimes during the COVID-19 pandemic (Findling *et al.*, 2022; Han *et al.*, 2023). Consequently, implementing automatic hate speech detection and moderation mechanisms is crucial to maintain the integrity of social media platforms and mitigate negative impacts in real-world scenarios, such as increased violence towards minorities (Laub, 2019).

Despite the well-established issue of online hate speech, numerous efforts have been made to detect and combat this problem (Schmidt and Wiegand, 2017; Del Vigna12 *et al.*, 2017). While state-of-the-art hate speech detection models have demonstrated

good performance on benchmark evaluation datasets, most of these models rely on transformer-based pre-trained language models or other deep neural network architectures (Sheth *et al.*, 2023) that lack interpretability and explainability. However, the task of hate speech detection is highly sensitive, and the interpretability of automated detectors is an essential and desirable feature. Sometimes, erroneous judgment may inadvertently strengthen the discrimination against the target group of the expression (Sap *et al.*, 2019; Davidson *et al.*, 2017). Thus, model interpretability is vital not only for end-user understanding but also for identifying biased predictions, and other prediction errors.

There are several ways to help with interpretability and explainability such as SHAP (SHapley Additive exPlanation) and LIME (Local Interpretable Model-agnostic Explanations). SHAP (Lundberg and Lee, 2017)) offers a unified framework for interpreting predictions of complex models by assigning importance values to each feature for a specific prediction. LIME (Ribeiro *et al.*, 2016)) addresses the opacity of machine learning models by explaining predictions in an interpretable and faithful manner through locally learned interpretable models around predictions. However, these interpretability metrics require significant computational resources for interpreting all data samples. Also, there is usually a trade-off between performance and model interpretability (Dziugaite *et al.*, 2020) which makes the interpretability difficult for complex models. Biased judgment-inducing expressions require contextual interpretation, akin to human judgment. Hence, hate speech detection models must contextualize and explain results for human comprehension (Kim *et al.*, 2022a).

Although incorporating interpretability directly into deep neural network models, such as pre-trained language model-based detectors, is challenging, one potential approach is to employ an auxiliary model to provide explanations or rationales, which are subsequently used in training the detection model. This method has been pro-

posed and utilized in the FRESH framework (Jain *et al.*, 2020), where the authors employ two separate networks: one for extracting task-specific rationales, and another that leverages those rationales to learn the classification task, thereby enabling faithful interpretability by construction. The FRESH (Jain *et al.*, 2020) paper focuses on providing faithful explanations for model predictions by simplifying the model architecture and using arbitrary feature importance scores for token selection. Nonetheless, the model's ultimate explanation is confined to tokens extracted by the method. In contrast, this thesis seeks to enhance interpretability in hate speech detection using Large Language Models (LLMs) to extract text-based rationales contributing to hatefulness. Thus, we propose a framework that utilizes large language models (LLMs) as the extractor model. We leverage the textual understanding and instruction-following capabilities of state-of-the-art LLMs to extract features from the input text, which are then used to augment the training of a separate base hate speech detector, facilitating faithful interpretability. The main contributions of this work are as follows:

1. We introduce **SHIELD**, a novel framework that utilizes rationales extracted by large language models (LLMs) to augment a base hate speech detection model, thereby enabling faithful interpretability.

2. We evaluate the quality of the features and rationales extracted by LLMs and measure their alignment with human-annotated rationales.

3. Through extensive experiments on datasets containing both implicit and explicit hate speech, we demonstrate that **SHIELD** maintains detection performance even after training with rationales for enhanced interpretability, overcoming the expected trade-off between interpretability and accuracy.

Chapter 2

RELATED WORKS

2.1 Hate Speech Detection

Hate speech detection is a well-established task that social media researchers have been dealing with in order to maintain the integrity of online platforms while upholding freedom of speech. In our modern world, where nearly everyone consumes social media data, it has become absolutely essential to mitigate the harmful effects caused by the proliferation of hate speech. While existing hate speech detection works have demonstrated remarkable capabilities, their complex nature often lacks transparency and interpretability. Existing complex deep learning models, while accurate, often operate as black boxes, making it challenging to explain their decision-making processes. Transparent and interpretable models would not only foster trust among end-users but also enable deeper insights into the workings of these systems, paving the way for more robust and equitable hate speech detection solutions that balance content moderation with the preservation of free speech.

Hate speech detection has garnered significant attention in recent years due to its impact on social media platforms and online communities. Various approaches and methodologies have been proposed to address the challenges associated with identifying and mitigating hate speech online. In this section, we provide an overview of relevant works in hate speech detection research, highlighting key contributions.

2.1.1 Traditional Hate Speech Detection Approaches

Several studies have indicated that offensive terms in text messages may be disguised through deliberate misspellings, typically involving the substitution of a single character (Warner and Hirschberg, 2012). The Levenshtein distance, which represents the minimal number of changes required to convert one string to another, offers a solution for this task. This distance metric can supplement dictionary-based methods effectively (Nandhini and Sheeba, 2015). Early research in hate speech detection focused on developing rule-based systems and handcrafted features to identify hate speech content. These systems relied heavily on predefined linguistic patterns and lexical cues to classify text as hateful or non-hateful. For instance, (Davidson *et al.*, 2017) introduced a dataset and classifier for hate speech detection on Twitter, utilizing handcrafted features and machine learning algorithms to distinguish between hate speech and other types of speech.

2.1.2 Machine Learning Approaches

Machine learning (ML) algorithms have made substantial contributions to the detection of hate speech and the analysis of social media content (Al-Garadi *et al.*, 2019). Offensive remarks such as hate speech (HS) and cyberbullying have received considerable attention in natural language processing (NLP) research over recent decades (Rodriguez *et al.*, 2019). ML algorithms have played a pivotal role in analyzing social media data to identify and categorize offensive comments (Weir *et al.*, 2018). The progress in ML algorithm research has had significant ramifications across various domains, resulting in the development of crucial tools and models for analyzing extensive datasets in real-world scenarios such as social media network (SMN) content analysis (Cheng *et al.*, 2015). With the advent of deep learning and natural

language processing techniques, machine learning approaches have become prevalent in hate speech detection research. These approaches leverage neural networks and advanced algorithms to automatically learn discriminative features from text data. For instance, (Nobata *et al.*, 2016) proposed a deep learning model for hate speech detection, employing convolutional neural networks (CNNs) to capture intricate patterns and semantics in text.

Recent advancements in multimodal learning have led to the exploration of combining textual and visual cues for hate speech detection. Multimodal approaches leverage both textual content and accompanying visual elements to enhance the understanding of hateful content. For example, (Zannettou *et al.*, 2019) proposed a multimodal framework for hate speech detection, incorporating both textual and visual features extracted from social media posts. Topic modeling is employed in identifying hateful remarks on major social media platforms such as YouTube (Latorre and Amores, 2021). In their study (Liu *et al.*, 2019a), the authors utilized the Latent Dirichlet Allocation model (Blei *et al.*, 2003) to uncover overarching themes and apply them in the classification of multimodal data.

2.1.3 Ethical Considerations and Bias Mitigation

As hate speech detection systems are deployed in real-world settings, ethical considerations and bias mitigation have become crucial aspects of research. Annotators' lack of sensitivity to variations in dialect may result in racial bias within automated hate speech detection models, possibly exacerbating harm against minority communities. Scholars have emphasized the importance of addressing biases in training data and algorithms to ensure fair and unbiased hate speech detection. For instance, the authors of the paper (Kiritchenko *et al.*, 2021) conducted a comprehensive review of NLP research on abusive content detection, with a specific emphasis on addressing

ethical challenges. They identified and emphasized the importance of eight ethical principles, ranging from privacy to fairness and human rights, to guide the development and deployment of hate speech detection technologies. Through this framework, they propose socio-technical solutions aimed at mitigating biases and potential harms associated with hate speech detection, including approaches like 'nudging', 'quarantining', and value-sensitive design, with the overarching goal of promoting rights-respecting and ethically sound practices throughout the technology's lifecycle. Similarly, in the work (Sap *et al.*, 2019), the authors conducted a comprehensive analysis of biases in hate speech detection datasets and proposed strategies to mitigate these biases.

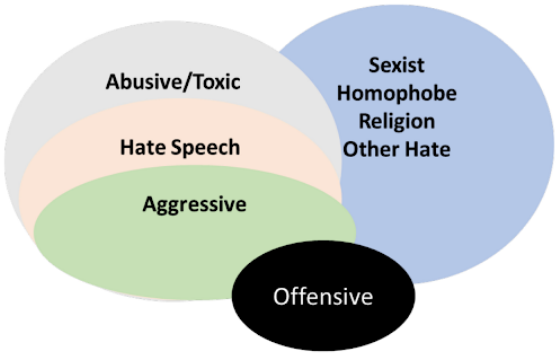

Figure 2.1: Hierarchy of Hate Speech Concepts (Alkomah and Ma, 2022)

Overall, there are two primary approaches to tackle hate speech detection. The first strategy involves leveraging additional or supplementary data sources. This includes utilizing user attributes (del Valle-Cano *et al.*, 2023), features of dataset annotators (Yin *et al.*, 2023), or understanding the consequences of hateful posts (Kim *et al.*, 2022b). For example, one study employed the implications of hateful posts to train a model on contrastive pairs representing hate content, aiming to detect implicit hate speech (Kim *et al.*, 2022b). Another study (Yin *et al.*, 2023) highlighted the challenge of achieving consensus among annotators on subjective tasks

like identifying hate speech and suggested incorporating definitive labels and annotator characteristics during training to enhance detection performance. A different study (del Valle-Cano *et al.*, 2023) analyzed users' social contexts and attributes to predict user satisfaction. However, the drawback of these methods is that accessing auxiliary information across different platforms is often difficult.

The second approach utilizes language models like BERT, which are pre-trained on large text datasets and renowned for their generalization capabilities. The performance of these models can be improved by fine-tuning them on specific hate speech datasets (Caselli *et al.*, 2021; Mathew *et al.*, 2021). One such example is HateBERT (Caselli *et al.*, 2021), a model fine-tuned on over 1.6 million hateful comments from Reddit, based on a BERT model. Similarly, HateXplain (Mathew *et al.*, 2021) is another model designed to detect and interpret hate speech. Other strategies include focusing on lexical cues (Schmidt and Wiegand, 2017) such as part-of-speech tags used (Markov *et al.*, 2021), facial expressions, content-related speech portions, or key phrases that convey hate (ElSherief *et al.*, 2018). To enhance language model representations, one study manually determined that sentiment and aggression are causal cues (Sheth *et al.*, 2023). Another study leveraged a causal graph to disentangle the input representations into platform-specific (hate-target-related features) and platform-invariant features to improve generalization capabilities for hate speech detection (Sheth *et al.*, 2024). Although effective, this method also requires auxiliary data (such as hate target labels), which are rarely available across various platforms.

Researchers have employed diverse techniques for hate speech detection and applied them to various problems. For instance, (Founta *et al.*, 2019) utilized Recurrent Neural Networks (RNNs) to classify racism and sexism. (Serra *et al.*, 2017) demonstrated the usefulness of character-level Long Short-Term Memory (LSTM) networks for detecting abusive language. Convolutional Neural Networks (CNNs) have also

proven successful in hate speech detection and classification tasks (Gambäck and Sikdar, 2017). More recently, large language models have been leveraged for these tasks, such as the work of (Wiedemann *et al.*, 2020), which proposes different fine-tuned and non-fine-tuned variations of pre-trained models like BERT (Devlin *et al.*, 2019), RoBERTa (Liu *et al.*, 2019b), and ALBERT (Lan *et al.*, 2019) for offensive language detection. The field of hate speech detection continues to evolve with advancements in machine learning, multimodal learning, and ethical considerations. While significant progress has been made, challenges such as bias mitigation and robustness against adversarial attacks remain areas of active research. By addressing these challenges, researchers aim to develop more effective and reliable hate speech detection systems for fostering safer online environments.

2.2 Large Language Models

Large Language Models (LLMs) are sophisticated artificial intelligence models capable of understanding and generating human-like text. These models are typically trained on vast amounts of text data and learn to predict the next word in a sequence based on the context provided by previous words. LLMs have gained significant attention due to their remarkable performance across various natural language processing (NLP) tasks. Transformer-based Language Models (LLMs) utilize an encoder-decoder architecture, wherein the encoder processes the input text into a series of contextualized representations, while the decoder generates the output text based on these representations. This architecture has shown remarkable effectiveness in various natural language processing tasks, including machine translation, text summarization, and question-answering. "Attention Is All You Need" by (Vaswani *et al.*, 2017) introduced the Transformer architecture, which laid the foundation for many subsequent large-scale language models.

In recent years, there has been a significant trend towards increasing the size of Language Models thereby improving their performance on various natural language understanding tasks. This trend, often referred to as "scaling up" LLMs, involves training models with larger numbers of parameters, which enables them to capture more complex patterns and nuances in language. For instance, the GPT-3 model developed by (Brown *et al.*, 2020) has 175 billion parameters, representing a substantial leap in model size compared to earlier iterations. Likewise, considerable effort has been dedicated to open-source Language Models (LLMs) like LLaMA (Touvron *et al.*, 2023) developed by Meta AI. This has introduced models spanning from 7B to 65B parameters, trained on vast volumes of data derived from publicly accessible datasets.

2.2.1 Feature Extraction with LLMs

Feature extraction involves capturing meaningful representations from input data, which can be utilized for downstream tasks such as classification, regression, or clustering. In the context of LLMs, feature extraction entails obtaining high-level representations of input text that capture semantic information. Large Language Models (LLMs) serve as powerful feature extractors by leveraging pre-trained representations learned from vast amounts of text data. Through techniques such as transfer learning and fine-tuning, LLMs can be effectively utilized across a wide range of downstream natural language processing tasks, showcasing their versatility and applicability in real-world scenarios.

In a recent work (He *et al.*, 2024a), the authors utilized Large Language Models (LLMs) as feature extractors which marked a significant advancement in natural language processing, particularly in the context of text-attributed graphs (TAGs). Traditional methods for handling text attributes in graphs often resort to simplistic representations, lacking the nuanced semantics captured by LLMs. By leveraging LLMs, which encode rich contextual information from vast corpora, as feature extractors, the model can effectively capture intricate textual nuances, enhancing the overall representation of nodes in the graph. This approach not only improves the interpretability of graph structures but also enables downstream tasks to leverage the comprehensive semantic understanding encoded within these features.

In a similar work that utilizes the LLM-guided Causal Explainability for Black-box Text Classifiers, the authors in the paper (Bhattacharjee *et al.*, 2024) explore the potential of Large Language Models (LLMs) in enhancing causal explainability for black-box text classification models. By leveraging the language understanding capabilities of LLMs, the authors propose a three-step pipeline to generate counterfactual

explanations, aiming to identify latent features influencing predictions and their associated input features. The results suggest promising prospects, particularly with high-quality LLMs like GPT-4, indicating the feasibility of using LLMs to enhance explainability in NLP tasks. This work opens avenues for further exploration of LLMs in causal explainability and broader applications in causal inference and discovery within the realm of NLP.

2.2.2 Significance of LLMs as Feature Extractors

With the recent advancements in large language models (LLMs) and their remarkable performance on various natural language processing tasks, researchers have started exploring their potential for enhancing interpretability and explainability in machine learning models. One promising approach is to leverage the language understanding capabilities of LLMs as feature extractors, providing explanatory features or rationales that can be used to augment and interpret the decisions of other models. In a related work (Baumann *et al.*, 2024), the authors explore the feasibility of employing large language models (LLMs), such as GPT-3.5 and GPT-4, for the extraction and normalization of attribute values from textual product titles and descriptions in e-commerce settings. They introduce the WDC Product Attribute-Value Extraction (WDC PAVE) dataset, which encompasses product offers from various websites and features manually verified attribute-value pairs. The methodology involves leveraging LLMs to process unstructured text, extracting attribute-value pairs, and then normalizing these values through operations like name expansion, generalization, and unit normalization. Their experiments demonstrate the effectiveness of LLMs, particularly GPT-4, in handling such tasks, showcasing their potential as feature extractors in e-commerce applications.

LLMs can be employed to extract salient features from textual data, aiding in the interpretation of model predictions. Researchers have explored techniques to utilize the internal representations of LLMs for feature extraction. For instance, (Yang *et al.*, 2024) demonstrated the effectiveness of using GPT-3 embeddings as features for interpreting clinical notes, showcasing the potential of LLMs in capturing meaningful representations from unstructured text data.

2.2.3 LLMs Understand Semantics

Interpretable machine learning has gained traction recently, and the rise of large language models (LLMs) presents an opportunity to redefine interpretability with a more ambitious scope across applications. Despite the potential for challenges such as hallucinated explanations and computational costs, their ability to articulate patterns in natural language offers insights understandable to humans. Thus, despite these limitations, the LLMs hold the potential to push interpretability boundaries. Two emerging research priorities are highlighted in (Singh *et al.*, 2024): using LLMs to directly analyze datasets and generate interactive explanations, leveraging their natural language capabilities to provide explanations at a complex scale while being human-interpretable. The authors advocate exploring LLMs' unique advantages to revolutionize the field of interpretable machine learning.

2.2.4 LLMs Understand Contextual Information

Information retrieval (IR) systems have evolved from traditional term-based methods to integrating advanced neural models, excelling at capturing contextual signals but facing challenges like data scarcity and generating inaccurate responses. The emergence of large language models (LLMs) like ChatGPT and GPT-4, with remarkable language understanding, generation, and reasoning abilities, has driven recent

research to leverage them for improving IR systems. In the survey (Zhu *et al.*, 2023a), the authors consolidate methodologies on using LLMs as query rewriters, retrievers, rerankers, and readers within IR systems, aiming to combine traditional term-based retrieval with LLMs' language understanding capabilities. It explores promising directions like search agents, providing a comprehensive overview of the confluence of LLMs and IR to address interpretability, data scarcity, and response accuracy challenges while enhancing retrieval performance through LLMs' language proficiency.

The utilization of LLMs as feature extractors has opened avenues for enhancing interpretability and explainability in various domains. By tapping into the rich representations learned by these models, researchers can extract meaningful features that facilitate the understanding of model predictions and decision-making processes.

Recent progress in Large Language Model (LLM) research has showcased enhanced performance not only in numerous natural language tasks (Min *et al.*, 2023) but also in more intricate domains like coding, mathematical reasoning, and others (Bubeck *et al.*, 2023). This advancement has spurred a line of inquiry aimed at assessing the efficacy of LLMs across diverse tasks. LLMs have demonstrated potential in tasks such as data annotation (He *et al.*, 2024b; Bansal and Sharma, 2023), information extraction (Dunn *et al.*, 2024), and even reasoning (Ho *et al.*, 2023). Due to their accessibility for querying, these models often serve as flawed experts or pseudo oracles in various tasks. Previous investigations have explored the feasibility of using language models as repositories of factual knowledge (Petroni *et al.*, 2019). In a recent study, researchers explored the application of LLMs in hate speech detection (Kumarage *et al.*, 2024). Authors in (Hasanain *et al.*, 2023) attempted propaganda span annotation using language models. However, our methodology differs in that we focus on leveraging the extracted spans, words, and rationales to enhance a detector model, thereby facilitating interpretability in an otherwise opaque model.

Chapter 3

FRAMEWORK

3.1 Motivation

Social media platforms have revolutionized the way people communicate and express themselves, providing a digital arena for interpersonal discussions and sharing of opinions (Bala, 2014; Nirmal *et al.*, 2023). However, the anonymity and perceived lack of consequences offered by these platforms have also enabled users to engage in the dissemination of hate speech and offensive content (Ullmann and Tomalin, 2020). As these platforms continue to grow in scale and reach, the need to automatically identify and flag instances of hate speech becomes increasingly crucial.

While several hate speech detection methods have been proposed, most of these approaches rely on complex deep learning models that operate as black boxes, lacking transparency and interpretability (Guidotti *et al.*, 2018). Interpretability is the degree to which a human can understand the cause of a decision (Miller, 2019). The lack of interpretability in these models raises concerns about their potential biases and other prediction errors. Interpretability is a critical requirement for hate speech detection systems, as it fosters trust among end-users and enables a deeper understanding of the decision-making process, ultimately leading to more robust and equitable solutions (Felzmann *et al.*, 2020).

To address this lack of interpretability, we propose a novel framework that leverages the capabilities of state-of-the-art Large Language Models (LLMs) to extract rationales, or explanatory features, from the input text. These rationales are then used to augment the training of a base hate speech classifier, enabling faithful interpretability

by design. Our approach effectively combines the textual understanding and reasoning abilities of LLMs with the discriminative power of state-of-the-art hate speech classifiers, resulting in models that are both accurate and inherently interpretable (Nirmal *et al.*, 2024).

Large Language Models (LLMs) have opened up a vast realm of research opportunities for AI researchers. Many believe that the era of LLMs will drive a boom in AI's ability to handle complex tasks with remarkable ease. These powerful models, known for fluently generating rich and contextualized text, are being explored for their potential to augment and enhance various AI applications (Hadi *et al.*, 2023). In the domain of misinformation detection, researchers are investigating how to leverage LLMs' generation capabilities to create more robust fake news detectors that can handle LLM-augmented disinformation more effectively (Jiang *et al.*, 2024). Additionally, LLMs have demonstrated prowess in several natural language tasks, such as stance detection for platform migration, closely aligning with human annotations (Jeong *et al.*, 2024). This has led researchers to explore leveraging the context understanding capabilities of LLMs to automate annotation tasks.

Following this line of thought, we aim to harness the power of LLMs to automate the process of obtaining rationales from human annotators for our use case. By leveraging LLMs in a one-shot manner, we seek to extract high-quality rationales while mitigating potential biases introduced by the models themselves. The goal is to capitalize on the strengths of LLMs, such as their language understanding and generation abilities, while maintaining the integrity and reliability of the final predictions. Through comprehensive evaluations on a diverse range of social media hate speech datasets, we demonstrate the efficacy of our framework in two key aspects: (1) the quality and alignment of the LLM-extracted rationales with human-annotated rationales, and (2) the surprising retention of detector performance even after train-

ing with rationales to ensure interpretability, defying the expected trade-off between interpretability and accuracy.

3.2 Model Architecture

We show our proposed **SHIELD** framework in Figure 3.1. In this section, we describe our framework in detail, elaborating on each of the components.

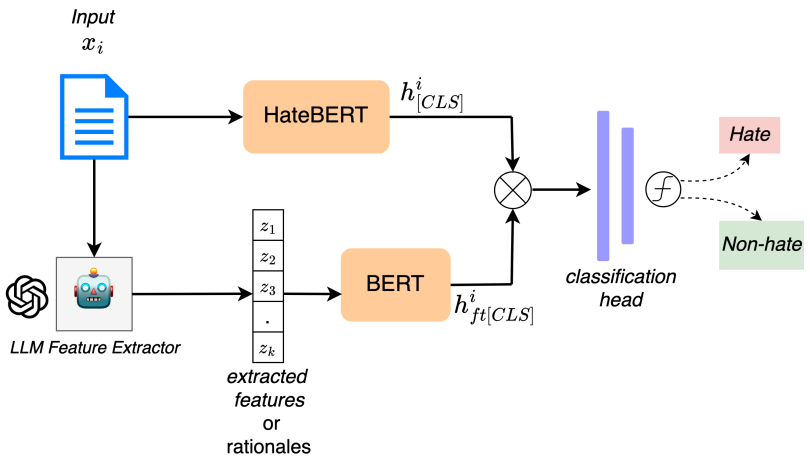

Figure 3.1: An Overview of the Proposed Framework Architecture.

3.2.1 LLM as Rationales Extractor

Our framework employs state-of-the-art instruction-tuned large language models (LLMs) as off-the-shelf textual feature extractors. While recent studies have demonstrated that LLMs struggle to perform hate speech detection tasks (Li *et al.*, 2023; Zhu *et al.*, 2023b) when used without additional models or fine-tuning, we hypothesize that we can leverage the textual understanding capabilities of these LLMs to extract textual features in the form of rationales. By restricting the use of LLMs to a simple text-level task, we aim to ensure that these models are not directly employed

for sensitive application tasks such as hate speech detection (Harrer, 2023), which could raise concerns about their potential biases or limitations.

Our approach seeks to harness the strengths of LLMs in comprehending and analyzing text while mitigating the risks associated with their direct application to sensitive tasks. By utilizing LLMs as auxiliary models for feature extraction, we can capitalize on their language understanding abilities while delegating the task of hate speech detection to a separate, dedicated model. This separation of concerns allows us to leverage the powerful capabilities of LLMs while maintaining the integrity and reliability of the hate speech detection process.

For a given input text $x_i \in X$, we utilize our carefully crafted task prompt to elicit the large language model (LLM) to extract features from the text that may convey hatefulness. In the context of explicit hate speech detection, such features could encompass categories like derogatory words, profanities, and other offensive language. Inspired by similar work (Bhattacharjee *et al.*, 2024), we also request the LLM to provide rationales explaining why the text should be classified as hateful or non-hateful. To perform this feature extraction process, for each input text, we prompt the LLM using the following prompt:

> "You are a content moderation bot. Identify the list of rationales, list of derogatory language, list of cuss words that promote a hateful sentiment and respond with non-hateful if there are none. Note: The output should be in a json format."
> Text: [input_text]

By leveraging the LLM's natural language understanding capabilities through a tailored prompt, we aim to extract relevant textual features and rationales that

can provide insights into the presence or absence of hateful content. The extracted features and rationales serve as supplementary inputs to a dedicated hate speech detection model, enabling it to make more informed and interpretable predictions. This approach allows us to capitalize on the LLM's strengths in text analysis while delegating the sensitive task of hate speech classification to a separate, specialized model.

3.2.2 Hate Speech Detector as Embedding Module

The next component in our framework is the base hate speech detector, which we aim to augment, such as HateBERT (Caselli *et al.*, 2021). HateBERT is a BERT model specifically fine-tuned on hate speech data. HateBert, an advanced model designed to identify hate speech, was created by refining a BERT (Devlin *et al.*, 2019) model using approximately 1.6 million hostile messages sourced from Reddit. For each input text $x_i \in X$, instead of obtaining the labels or class probabilities, we extract the last layer embedding of the `[CLS]` token, $h^i_{[CLS]}$, which essentially encapsulates all the relevant information from the input text for the hate speech detection task.

By leveraging the pre-trained and fine-tuned representations learned by HateBERT, our framework gains access to a rich and task-specific encoding of the input text. Rather than relying solely on the final classification output, we utilize the `[CLS]` token embedding, which serves as a condensed representation of the input text's semantics and features pertinent to hate speech detection. This approach allows us to augment the base hate speech detector with additional features and rationales extracted by the large language model, enabling the creation of a more interpretable hate speech detection system.

3.2.3 Feature Embedding Model

After post-processing the outputs, we have a list of k textual features $\{z_j\}_{j=1}^{k}$ for the given input text x_i. For these textual features and rationales extracted via the large language model (LLM), we employ a pre-trained transformer-based language model (PLM), such as BERT, to embed these features. PLMs, even without any task-specific fine-tuning, provide rich, expressive latent representations for text. Therefore, we feed the LLM-extracted textual features into a BERT model (specifically, bert-base-uncased [1]) and obtain the last hidden layer embedding of the `[CLS]` token, which we denote as $h^i_{ft[CLS]}$.

Thus, leveraging a pre-trained language model like BERT, we can obtain informative and contextualized representations of the LLM-extracted textual features and rationales. These embeddings, denoted as $h^i_{ft[CLS]}$, capture the semantic and contextual information present in the LLM-extracted features, enabling us to incorporate them effectively into our hate speech detection framework. The rich representations provided by PLMs, even without task-specific fine-tuning, allow us to leverage their language understanding capabilities and augment the base hate speech detector with complementary information from the LLM-extracted features and rationales.

3.2.4 Embedding Fusion & Classification

For each input text x_i, we have obtained two embeddings from the previous components: the text embedding $h^i_{[CLS]}$ from the base hate speech detector, and the feature embedding $h^i_{ft[CLS]}$ from the feature embedding BERT model. To combine these two embeddings, we simply concatenate them:

[1] https://huggingface.co/google-bert/bert-base-uncased

$$h^i_{combined} = h^i_{[CLS]} \oplus h^i_{ft[CLS]} \tag{3.1}$$

By concatenating the text embedding from the base hate speech detector and the feature embedding from the LLM-extracted textual features and rationales, we create a comprehensive representation, $h^i_{combined}$, that encapsulates information from both sources. This combined embedding serves as a rich and augmented input for the final hate speech classification task, incorporating the task-specific knowledge from the base detector and the complementary textual features and rationales provided by the large language model.

The concatenation operation allows us to fuse the two embeddings seamlessly, preserving the individual information from each component while enabling the final classification model to leverage the combined representation effectively. This approach facilitates the integration of the LLM-extracted features and rationales into the hate speech detection pipeline, ultimately enhancing the model's interpretability and decision-making capabilities.

It is important to note that while the authors in (Jain *et al.*, 2020) only utilized the extracted rationales in the subsequent detector model, we employ a concatenated view to incorporate additional contextual features that may be highly relevant for determining the hate or non-hate label (Ocampo *et al.*, 2023). We feed this combined embedding $h^i_{combined}$ into a feed-forward multi-layer perceptron with two fully connected layers and a ReLU activation (Agarap, 2018) in between, to project it onto a smaller dimension space. Following previous work (Pan *et al.*, 2022; Bhattacharjee *et al.*, 2023), we adopt this approach to retain important features and avoid overfitting the model during training. We denote this MLP as $f(\cdot)$.

Finally, we compute the batch-wise binary cross-entropy loss using the ground truth label y_i for each input text x_i:

$$loss_{CE} = -\frac{1}{n}\sum_{i}^{n}[(y_i|f(h^i_{combined})) + (1-y_i)\log(1-p(y_i|f(h^i_{combined}))] \tag{3.2}$$

where n is the batch size. Since we are using the BERT feature embedding model solely to encode the textual features z, we keep this model frozen and train the remainder of the framework with this simple loss.

Here we see that by concatenating the text embedding from the base hate speech detector and the feature embedding from the LLM-extracted textual features and rationales, we create a comprehensive representation that incorporates information from both sources. This combined embedding serves as a rich input for the final hate speech classification task, enhancing the model's ability to leverage complementary information and make more informed decisions. The subsequent multi-layer perceptron projects thus combined embedding onto a smaller dimension space, retaining important features while mitigating overfitting during training.

3.3 Datasets

To assess the effectiveness of our proposed **SHIELD** framework, we utilize a combination of explicit and implicit hate speech datasets sourced from various social media platforms, all of which are in the English language. For explicit hate speech, we incorporate publicly available benchmark datasets from prominent platforms including GAB, Twitter, YouTube, and Reddit. The **GAB** dataset (Mathew *et al.*, 2021) comprises annotated posts from the GAB website, with binary labels indicating the presence of hateful content. Similarly, the **Reddit** dataset (Kennedy *et al.*, 2020) consists of posts labeled as either hateful or non-hateful. The **Twitter** dataset (Mathew *et al.*, 2021) contains instances of hate speech extracted from tweets on the Twitter platform. Lastly, the **YouTube** dataset (Salminen *et al.*, 2018) comprises

expressions and comments deemed hateful on the YouTube platform. To ensure consistency, we preprocess these datasets following the methodology outlined in (Sheth *et al.*, 2024), resulting in cleaned binary labels. Table 3.1 provides a summary of the datasets, including the distribution of hateful and non-hateful posts.

In addition to explicit forms of hate speech, we also consider implicit instances in our assessment. While subtle expressions of abuse may not immediately appear overtly harmful, their covert nature can lead to comparable levels of harm over time. Hence, detecting implicit hate speech becomes increasingly crucial. To address this, we assess the performance of our proposed model using the **Implicit Hate Speech Corpus** (ElSherief *et al.*, 2018). This corpus consists of Twitter posts annotated as explicit hate, implicit hate, or non-hate speech. We focus exclusively on implicit hate and non-hate instances for our binary classification task.

Dataset	**# of Posts**	**# of Hateful Posts**	**Hate %**
GAB	14,240	11,920	83.7
Reddit	37,164	10,562	28.4
Twitter	10,457	3,933	37.6
YouTube	5,052	1,699	33.6
Implicit	20,391	7,100	34.8

Table 3.1: Dataset Statistics for Hate Speech Datasets from Different Social Media Platforms.

3.4 Baselines

We compare our proposed **SHIELD** framework to a variety of different baselines in order to understand the impact of the augmentation with rationales. We use the following well-known baseline hate speech detection models:

HateBERT: This is also the base model used in our framework. HateBERT (Caselli *et al.*, 2021) uses over 1.5 million Reddit messages from suspended communities known for encouraging hate speech to fine-tune the BERT-base model. We further fine-tune HateBERT on each dataset and report the performance.

HateXplain: Similarly, we fine-tune the HateXplain (Mathew *et al.*, 2021) model on each of our datasets and report the performance. HateXplain model is trained on hateful posts along with the target community, the rationales, and the portion of the post on which human annotators' labelling decision is based.

PEACE: We further extend our comparison on PEACE (Sheth *et al.*, 2023) framework which uses Sentiment and Aggression Cues to detect the overall sentiment of the text.

CATCH: Furthermore, we compare our model with CATCH (Sheth *et al.*, 2024) framework which disentangles the input representations into invariant and platform-dependent features.

ChatGPT-1shot: Apart from these hate speech specific detection models, we also compare our framework with an off-the-shelf GPT-3.5 model, to understand how well the LLM performs on the same datasets. We do this in a one-shot manner, i.e., by providing the task instruction along with an example input and ground truth label.

3.5 Experiments

To implement our proposed **SHIELD** framework, we leverage PyTorch along with the Huggingface Transformers library as illustrated in Figure 3.1, our initial component employs a readily available Large Language Model (LLM) to extract features and rationales. Specifically, we utilize OpenAI's GPT-3.5 (specifically, GPT-3.5-turbo-0613) [2] , which has demonstrated remarkable performance across various Natural Language Processing (NLP) tasks (Guo *et al.*, 2023). Access to this model is facilitated through the OpenAI API. To extract features and generate rationales, we configure the temperature to 0.1 and top_p to 1. In addition, we utilize a pre-trained, frozen BERT (bert-base-uncased) for the Feature Embedding Model and a pre-trained HateBERT model for the Hate Speech Detector [3] . We employ the AdamW optimizer with a learning rate of 2×10^{-5}. Model training is conducted on two machines: one equipped with an NVIDIA GP102 [TITAN Xp] GPU with 12 GB VRAM, and another featuring an NVIDIA A100 GPU with 40GB RAM. Throughout all the experiments, we assess performance using accuracy as the evaluation metric.

[2] or otherwise commonly referred to as 'ChatGPT'

[3] https://huggingface.co/GroNLP/hateBERT

Chapter 4

RESULTS

In this section, we detail our experiments and provide an in-depth analysis of the experimental findings. Our objective is to investigate the viability and efficacy of our proposed **SHIELD** framework by addressing the following research questions:

- **RQ1:** How effective is ChatGPT's performance on our collection of hate speech detection datasets?
- **RQ2:** Can we leverage the capabilities of recent state-of-the-art large language models (LLMs) to extract relevant features in the form of rationales, and do these rationales align with human judgment?
- **RQ3:** Can our proposed **SHIELD** framework effectively maintain or improve the performance of the hate speech detector while simultaneously facilitating faithful interpretability?

In the first research question, we aim to evaluate the performance of ChatGPT, a prominent language model, on our curated set of hate speech detection datasets. This assessment will provide a baseline understanding of the model's capabilities in this specific task.

The second research question explores the potential of leveraging state-of-the-art LLMs as feature extractors, specifically focusing on their ability to extract rationales that can provide insights into the presence or absence of hate speech. Additionally, we seek to determine whether these LLM-extracted rationales align with hu-

man judgment, which is crucial for ensuring the interpretability and reliability of our framework.

Lastly, the third research question investigates the effectiveness of our proposed **SHIELD** framework in retaining or improving the performance of the hate speech detector while enabling faithful interpretability. This aspect is critical as it addresses the trade-off between model performance and interpretability, aiming to achieve both objectives simultaneously.

4.1 Performance of ChatGPT on the hate speech detection task

Several recent studies investigate whether Large Language Models have the capability to replicate human-annotated ground truth labels in social computing tasks(Zhu *et al.*, 2023b). However, despite extensive pre-training on vast datasets, where Large Language Models are anticipated to excel in this task, this is not always the case. To further scrutinize this matter beyond the scope of prior research, we meticulously devise a one-shot prompt and prompt ChatGPT to classify the input text based on a labeled example in the prompt. The result of this prompt yields a single label, with hateful text represented as label "1" and non-hateful text represented as label "0" as depicted in Table 4.1.

We conduct this classification task using ChatGPT across all five datasets and calculate the accuracy. Subsequently, we compare the outcomes of this one-shot classification task with those of the baseline Models (as outlined in Section 3.4), presenting the findings in Table 4.2. A notable discrepancy is observed between the performance of the baseline models and the accuracy of ChatGPT-1shot classification. While the performance on the GAB dataset appears satisfactory, ChatGPT encounters challenges with the remaining four datasets, achieving approximately around 58% to 65%. Similar findings have been documented in recent research exploring the out-of-

Text	"always thought it astounding no policing agency dares do this to the armed muslim compounds all over the country"
Prompt	*You are a hate speech detection bot. Given a text, respond with the classification label towards given text with either hateful labeled as 1 or non-hateful labeled as 0. Note: It is essential to give classification for all the texts.* *Text:* "always thought it astounding no policing agency dares do this to the armed muslim compounds all over the country"
ChatGPT response	1

Table 4.1: Examples of Input Text, Prompt, and ChatGPT's Response for a Data Sample from the Twitter Dataset.

the-box performance of Large Language Models (LLMs) in hate speech detection (Li *et al.*, 2023; Zhu *et al.*, 2023b).

Although this highlights the difficulty ChatGPT and potentially other Large Language Models (LLMs) face in directly detecting hate speech, these models have demonstrated remarkable comprehension of textual content. Utilizing these models solely for extracting features or rationales, rather than for conducting the entire detection task, could prove advantageous.

	HateBERT	HateXplain	PEACE	CATCH	ChatGPT-1shot	SHIELD (ours)
GAB	96	**97**	91	82	85.39	<u>96.3</u>
YouTube	71	<u>72</u>	<u>72</u>	**79**	58.34	70
REDDIT	<u>94</u>	93	93	86	65.05	**94.5**
Twitter	56	60	31	**78**	60.09	<u>64</u>
Implicit HS	**78**	<u>76</u>	64	–	65.68	**78**

Table 4.2: Evaluation Results for Our Shield Framework Vs. The Baseline Models. Implicit HS Refers to the Implicit Hate Speech Corpus. Values in **Bold** Denote the Best Performance, and <u>Underlined</u> Values Denotes the Second-best Performance.

4.2 Goodness of ChatGPT extracted features or rationales

We aim to assess ChatGPT's capacity for understanding both text and context, with a focus on extracting features that are relevant to hate speech detection. Following a methodology similar to that outlined in previous research (Jain *et al.*, 2020), we employ the Large Language Model (LLM), specifically GPT-3.5, as the *extractor* model. Unlike the extractor model described in prior work (Jain *et al.*, 2020), our approach does not necessitate additional fine-tuning for task-specific purposes, owing to the advanced instruction-following capabilities of modern LLMs. We design a prompt (outlined in Table 4.3) to extract *rationales*, *derogatory language* and, *cuss words* from input text, serving as interpretable features for subsequent use in the *predictor* model (HateBERT), thus facilitating a faithful interpretation of hate speech detection results. To evaluate the quality of the extracted features or rationales, we compare ChatGPT-extracted rationales with human-annotated ground truth rationales using the annotated rationale spans provided in the HateXplain dataset (Mathew *et al.*, 2021). Following standard preprocessing steps such as stop word removal, we compute

Prompt: *You are a content moderation bot. Identify the list of rationales, list of derogatory language, list of cuss words that promote a hateful sentiment and respond with non-hateful if there are none. Note: The output should be in a json format. "Text"*		
Dataset	**Text**	**ChatGPT-extracted features and rationales**
GAB	"does this mean they arrested the mayor of london have cordoned off every muzrat ghetto in the uk every mosque closed down if not its business as usual"	arrested the mayor of london, cordoned off every muzrat ghetto in the uk, every mosque closed down; muzrat, ghetto
REDDIT	"Shut up! 50 years ago we'd have you upside down with a f**king fork up your a**! You can talk, you can talk, you can talk! You're brave now motherf**ker! Throw his a** out, he's a n**ger! HE'S A N**GER! HE'S A N**GER! A N**GER! LOOK, THERE'S A N**GER!"	n**ger; f**king, motherf**ker
Twitter	"show me the black muslims bc im calling bs on this and you should too"	black muslims; black, muslims
YouTube	"Dirty f**king israeli zionist jews"	israeli, zionist, jews; f**king
Implicit HS	"i'm so glad #ferguson happened . you woke the world up to how much black people hate white people ."	ferguson, black people hate white people; black people, white people

Table 4.3: Examples of Input Text along with the LLM-extracted Features and Rationales. Rationales Are in Blue, Derogatory Language Is in Red, Cuss Words Are in Teal.

the similarity between the ChatGPT-extracted rationales and the human-annotated rationales for input text from the HateXplain dataset, reporting these scores in Table 4.4. Our analysis reveals a significant overlap in token space and a high level of semantic similarity in the latent space between the rationales produced by the LLM and those annotated by humans.

We showcase examples from each of the five datasets in Table 4.3, which include the input text labeled as 'hateful' along with the features extracted by ChatGPT. The features are categorized into three distinct groups, highlighted in different colors: rationales, derogatory language, and cuss words. It's evident that the LLM effectively identifies these words and spans with precision.

We also provide examples in Figure 4.1 to visually compare the human-annotated

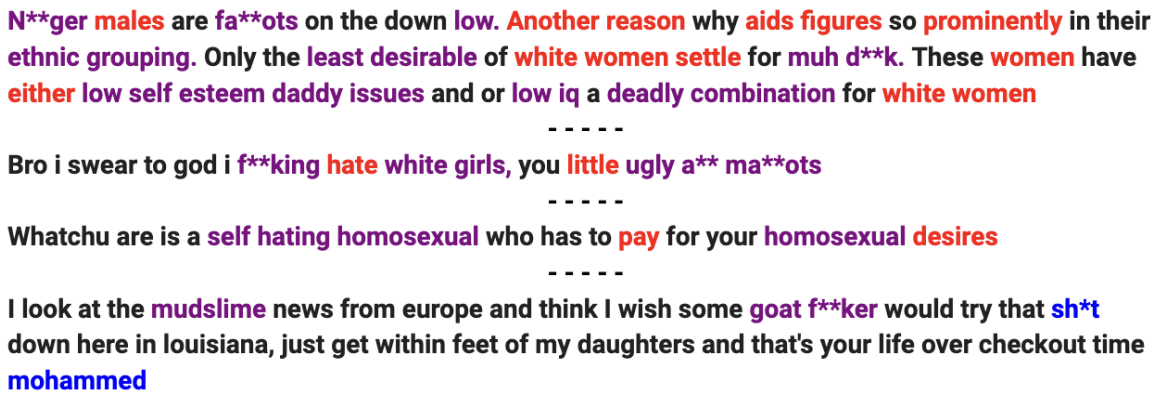

Figure 4.1: Examples with Both LLM-Annotated and Human-Annotated Rationales. Overlap Is in Purple.

rationales with those extracted by the LLM. Text highlighted in red represents rationales annotated by human annotators, while text highlighted in blue denotes rationales or words identified by the LLM. Spans highlighted in purple indicate areas where both the LLM and human annotations overlap. From these examples, it's evident that there is a significant degree of overlap, indicating that the LLM effectively captures semantically relevant segments of the text. Interestingly, we observe that while human annotators sometimes annotate words or spans with less relevance to the task, the LLM-extracted rationales do not include these spans (such as 'aids figures' and 'prominently' in the first example in Figure 4.1). Leveraging LLM-extracted rationales for training could prove even more beneficial in such scenarios, as it allows for the avoidance of some of the noisy signals present in the data.

Similarity Metric	Similarity Coefficients
Jaccard similarity	60.39%
Overlap Similarity	99.17%
Cosine Similarity	74.51%
Google's Universal Sentence Encoder	56.09%

Table 4.4: Similarity Between HateXplain Human Explanations and LLM-extracted Features/Rationales.

4.3 Hate speech detector performance after training with extracted rationales

In this experiment, we aim to train a hate speech detector by incorporating the extracted rationales into the input text, facilitating faithful interpretability of the classifications. For this purpose, we utilize a HateBERT model as the base hate speech detector model and report the results in Table 4.2, along with the results from other baselines. We observe that our proposed **SHIELD** framework performs on par with a HateBERT model fine-tuned on the same dataset, i.e., at par with the base model. This performance retention is encouraging, as models are otherwise known to trade-off accuracy for interpretability (Dziugaite *et al.*, 2020; Bersimas *et al.*, 2019). Interestingly, in the Twitter dataset, we see a significant 12.5% performance jump by our **SHIELD** model compared to the fine-tuned HateBERT model. This potential improvement might be due to noise in the Twitter dataset: the extracted rationales may provide more discriminative training signals, allowing the detector to train on robust features instead of noisy ones. However, further analysis is required to verify this claim.

Thus, by incorporating the extracted rationales into the input text, our **SHIELD** framework aims to facilitate faithful interpretability of the hate speech detector's classifications. The fact that our framework achieves performance on par with the

base model, while enabling interpretability, is a promising result, as there is often a trade-off between accuracy and interpretability in models. Furthermore, the significant performance improvement observed in the Twitter dataset may suggest that the extracted rationales possibly help mitigate the impact of noise in the data, allowing the detector to focus on more discriminative features. While this claim requires further investigation, the initial results demonstrate the potential of our **SHIELD** framework in delivering both accurate and interpretable hate speech detection.

4.4 Interpretability with respect to human annotations

	HateBERT	**HateXplain**	**PEACE**	**SHIELD (ours)**
Jaccard Similarity	48.18	48.12	43.00	**60.39**
Overlap Similarity	98.94	99.05	94.85	**99.17**
Cosine Similarity	63.42	63.36	63.13	**74.51**
Google's Universal Sentence Encoder	38.57	38.81	27.09	**56.09**

Table 4.5: Interpretability Metrics: Similarity Between Human Annotated Rationales and Model Attention Tokens rationales for the HateXplain dataset

We conduct additional experiments to examine the similarity between the rationales generated by the Large Language Model and those annotated by humans using the HateXplain dataset, which is a readily available dataset of human-annotated data on hate speech. Furthermore, we also measure the similarities between these human annotated rationales with that of rationales inferred from model's attention scores for each input. We extract the attention scores of the input tokens for each input text from each of the models specified in Table 4.5, and we compute the similarity metrics in both token space and latent space. Employing standard preprocessing tech-

niques including stop word removal, we compute the similarity between ChatGPT-extracted rationales and human-annotated rationales for input text sourced from the HateXplain dataset, as depicted in Table 4.5. Our analysis underscores a notable convergence and substantial semantic similarity between the rationales generated by the LLM and those annotated by humans. Notably, we observe that the similarity metrics between human-generated tokens and LLM-generated tokens surpass the similarities between those of human-annotated rationales and model attention tokens, as depicted in Table 4.5. This signifies the superior alignment between human and LLM-generated rationales, emphasizing the faithfully interpretable nature of our proposed SHIELD framework.

4.5 Modifying the hate speech detector and feature embedding models

In order to gain additional insights into the impact of the framework components, we modify the choice of the base pre-trained language models used for the hate speech detector and the feature extractor. The specific variations we experiment with are: (1) the original **SHIELD** framework, which employs HateBERT as the hate speech detector (HSD) and bert-base-uncased as the feature embedding model (FE), (2) **SHIELD** with a pre-trained roberta-base as the HSD instead of HateBERT, and (3) **SHIELD** with a pre-trained roberta-base as the FE instead of bert-base-uncased. We choose to perform this analysis with RoBERTa (Liu *et al.*, 2019b) instead of the two BERT-based models, as RoBERTa has been shown to sometimes outperform BERT (Devlin *et al.*, 2019) on various natural language understanding tasks (Tarunesh *et al.*, 2021).

We report the results of this analysis in Table 4.6. Overall, we observe some variation in performance based on the model choice for the HSD and FE components. While using roberta-base as the FE component marginally improves performance for

only one dataset, i.e., GAB, employing roberta-base as the HSD instead of HateBERT achieves higher performance for three datasets. This is particularly interesting since, unlike HateBERT, the pre-trained roberta-base is not specifically trained on the hate speech task.

Here, by modifying the base pre-trained language models used for the hate speech detector and the feature extractor, we aim to analyze the impact of these components on the overall performance of our **SHIELD** framework. The variations in performance observed across different model choices provide insights into the framework's flexibility and the potential benefits of exploring alternative pre-trained models. Notably, the improved performance achieved with roberta-base as the HSD, despite not being specifically trained for hate speech detection, highlights the framework's ability to leverage the strengths of different pre-trained models effectively.

Overall, **SHIELD** shows promising results in leveraging LLM-extracted rationales into augmenting a base hate speech detector, to facilitate faithful interpretability, while maintaining detection performance. We confirm that the rationales extracted by the LLM are consistent with human judgment. Our framework undergoes training and evaluation across various benchmark datasets containing both implicit and explicit hate speech sourced from diverse online social media platforms. We illustrate how our **SHIELD** framework manages to uphold performance levels akin to the base model, even amidst an anticipated trade-off between accuracy and interpretability. Consequently, we present a hate speech detection system that remains faithfully interpretable, leveraging LLM-extracted rationales instead of human annotations.

	GAB	YouTube	REDDIT	Twitter	Implicit HS
SHIELD (roberta-base HSD)	87.53	**72.2**	84.8	**67.03**	**78.36**
SHIELD (roberta-base FE)	**96.42**	69.27	94.21	56.22	77.52
SHIELD	96.3	70	**94.5**	64	78

Table 4.6: Analysis of HSD and FE Model Choices in the SHIELD Framework. HSD: Hate Speech Detector, FE: Feature Embedding Model. The Original SHIELD Framework Has Hatebert as the Hate Speech Detector and Bert-base-uncased as the Feature Embedding Model. Numbers in Bold Denote Best Performing Model Variant for Each Dataset.

Chapter 5

CONCLUSION AND FUTURE WORK

5.1 Summary

In this work, we explore the challenge of hate speech detection on social media platforms and propose a method to train interpretable classifiers using rationales extracted by large language models (LLMs). Recognizing the unsatisfactory performance of LLMs as detectors for hate speech, we instead aim to leverage the textual understanding and instruction-following capabilities of LLMs, such as ChatGPT, to extract words and rationales from the text that are associated with the hate speech label. We propose a framework called **SHIELD**, which utilizes these LLM-extracted rationales to augment the training of a base hate speech detector, facilitating faithful interpretability. We verify the alignment of the LLM-extracted rationales with human judgment. We train and evaluate our framework on multiple benchmark datasets comprising both implicit and explicit hate speech from various online social media platforms.

Through our comprehensive evaluation, we demonstrate how our **SHIELD** framework can maintain performance similar to the base model, despite the expected trade-off between accuracy and interpretability. Consequently, we introduce a faithfully interpretable hate speech detector that relies solely on LLM-extracted rationales instead of human-annotated rationales. By harnessing the capabilities of LLMs to extract relevant rationales and augmenting the training process of a base hate speech detector, our **SHIELD** framework addresses the critical need for interpretability in hate speech detection systems. This approach not only facilitates the development

of interpretable models but also maintains their performance, offering a promising solution to the challenge of balancing accuracy and interpretability in this sensitive and consequential task.

5.2 Future Work

Although our work follows a similar approach to Jain *et al.* (2020) and we establish faithfulness by construction, future research could explore more robust ways to evaluate the faithfulness of the resulting hate speech detector. In this study, we verified the quality of the extracted rationales by comparing them with the ground truth for one dataset. However, future work can investigate better automated methods to evaluate and verify the quality of the LLM-extracted rationales. Future research could focus on developing more sophisticated and automated techniques to comprehensively evaluate and verify the quality of the rationales extracted by large language models. Such methods could potentially involve automated metrics, comparative analyses, or other data-driven approaches to assess the alignment and faithfulness of the extracted rationales with respect to the underlying task and domain-specific nuances.

By exploring better ways to evaluate the faithfulness of the resulting detector, researchers can further enhance the reliability and trustworthiness of the proposed approach, ensuring that the interpretability facilitated by the LLM-extracted rationales is truly faithful and accurately reflects the decision-making process of the hate speech detector. This line of investigation could contribute to the development of more robust and reliable interpretable AI systems, particularly in sensitive domains like hate speech detection, where transparency and accountability are paramount.

5.3 Limitations

While our **SHIELD** framework shows promise in leveraging large language models (LLMs) to create interpretable hate speech detectors, several limitations need to be addressed. In certain cases, the LLM may fail to identify coherent rationales, leading to incomplete or inaccurate explanations for the model's predictions. The choice of the LLM itself is also crucial, as powerful proprietary models like ChatGPT may not be accessible to all researchers, while open-source alternatives could potentially yield suboptimal performance. Our current work utilizes ChatGPT for rationale extraction, but exploring the capabilities of different LLMs, including multilingual and domain-specific models, could provide valuable insights. Additionally, our framework may need adaptation to handle instances where the LLM cannot provide clear rationales, either through ensemble methods or by incorporating human feedback mechanisms to refine the extracted rationales. This could involve combining rationales from multiple LLMs or allowing human experts to review and refine the rationales, ensuring more accurate and reliable explanations.

Furthermore, since these LLMs lack the ability to fully grasp context independently, they introduce some level of noise. It may be beneficial to explore denoising methods in the future. Despite the absence of denoising, we have noticed enhanced performance in certain datasets like Twitter. However, the considerable noise makes it challenging to capture specific rationales, leading to suboptimal performance gains in other datasets. Therefore, in the future, employing denoising techniques could lead to better context understanding and improved performance in hate speech detection classification tasks.

REFERENCES

Agarap, A. F., "Deep learning using rectified linear units (relu)", Neural and Evolutionary Computing (2018).

Al-Garadi, M. A., M. R. Hussain, N. Khan, G. Murtaza, H. F. Nweke, I. Ali, G. Mujtaba, H. Chiroma, H. A. Khattak and A. Gani, "Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges", IEEE Access **7**, 70701–70718 (2019).

Alkomah, F. and X. Ma, "A literature review of textual hate speech detection methods and datasets", Information **13**, 6, 273 (2022).

Bala, K., "Social media and changing communication patterns.", Global Media Journal: Indian Edition **5**, 1 (2014).

Bansal, P. and A. Sharma, "Large language models as annotators: Enhancing generalization of nlp models at minimal cost", arXiv preprint arXiv:2306.15766 (2023).

Baumann, N., A. Brinkmann and C. Bizer, "Using llms for the extraction and normalization of product attribute values", arXiv preprint arXiv:2403.02130 (2024).

Bersimas, D., A. Delarue, P. Jaillet and S. Martin, "The price of interpretability", (2019).

Bhattacharjee, A., T. Kumarage, R. Moraffah and H. Liu, "Conda: Contrastive domain adaptation for ai-generated text detection", in "Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)", pp. 598–610 (2023).

Bhattacharjee, A., R. Moraffah, J. Garland and H. Liu, "Towards llm-guided causal explainability for black-box text classifiers", in "AAAI 2024 Workshop on Responsible Language Models, Vancouver, BC, Canada", (2024).

Blei, D. M., A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation", Journal of machine Learning research **3**, Jan, 993–1022 (2003).

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners", Advances in neural information processing systems **33**, 1877–1901 (2020).

Bubeck, S., V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg *et al.*, "Sparks of artificial general intelligence: Early experiments with gpt-4", arXiv preprint arXiv:2303.12712 (2023).

Caselli, T., V. Basile, J. Mitrović and M. Granitzer, "Hatebert: Retraining bert for abusive language detection in english", in "Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)", pp. 17–25 (2021).

Cheng, J., C. Danescu-Niculescu-Mizil and J. Leskovec, "Antisocial behavior in online discussion communities", in "Proceedings of the international aaai conference on web and social media", vol. 9, pp. 61–70 (2015).

Davidson, T., D. Warmsley, M. Macy and I. Weber, "Automated hate speech detection and the problem of offensive language", in "Proceedings of the international AAAI conference on web and social media", vol. 11, pp. 512–515 (2017).

del Valle-Cano, G., L. Quijano-Sánchez, F. Liberatore and J. Gómez, "Socialhaterbert: A dichotomous approach for automatically detecting hate speech on twitter through textual analysis and user profiles", Expert Systems with Applications **216**, 119446 (2023).

Del Vigna12, F., A. Cimino23, F. Dell'Orletta, M. Petrocchi and M. Tesconi, "Hate me, hate me not: Hate speech detection on facebook", in "Proceedings of the first Italian conference on cybersecurity (ITASEC17)", pp. 86–95 (2017).

Devlin, J., M.-W. Chang, K. Lee and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", in "Proceedings of NAACL-HLT", pp. 4171–4186 (2019).

Dunn, A., J. Dagdelen, N. Walker, S. Lee, A. S. Rosen, G. Ceder, K. Persson and A. Jain, "Structured information extraction from complex scientific text with fine-tuned large language models", Nature communications (15 Feb. 2024).

Dziugaite, G. K., S. Ben-David and D. M. Roy, "Enforcing interpretability and its statistical impacts: Trade-offs between accuracy and interpretability", ArXiv **abs/2010.13764**, URL `https://api.semanticscholar.org/CorpusID:225075958` (2020).

ElSherief, M., 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 International AAAI Conference on Web and Social MediaProceedings of the International AAAI Conference on Web and Social Media", vol. 12 (2018).

Felzmann, H., E. Fosch-Villaronga, C. Lutz and A. Tamò-Larrieux, "Towards transparency by design for artificial intelligence", Science and engineering ethics **26**, 6, 3333–3361 (2020).

Findling, M. G., R. J. Blendon, J. Benson and H. Koh, "Covid-19 has driven racism and violence against asian americans: perspectives from 12 national polls", Health Affairs Forefront (2022).

Founta, A. M., D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali and I. Leontiadis, "A unified deep learning architecture for abuse detection", in "Proceedings of the 10th ACM conference on web science", pp. 105–114 (2019).

Gambäck, B. and U. K. Sikdar, "Using convolutional neural networks to classify hate-speech", in "Proceedings of the first workshop on abusive language online", pp. 85–90 (2017).

Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, F. Giannotti and D. Pedreschi, "A survey of methods for explaining black box models", ACM computing surveys (CSUR) **51**, 5, 1–42 (2018).

Guo, K., 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 "2023 International Conference on Machine Learning and Applications (ICMLA)", pp. 1568–1573 (IEEE, 2023).

Hadi, M. U., R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili *et al.*, "A survey on large language models: Applications, challenges, limitations, and practical usage", Authorea Preprints (2023).

Han, S., J. R. Riddell and A. R. Piquero, "Anti-asian american hate crimes spike during the early stages of the covid-19 pandemic", Journal of interpersonal violence **38**, 3-4, 3513–3533 (2023).

Harrer, S., "Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine", EBioMedicine **90** (2023).

Hasanain, M., F. Ahmed and F. Alam, "Large language models for propaganda span annotation", arXiv preprint arXiv:2311.09812 (2023).

He, X., X. Bresson, T. Laurent and B. Hooi, "Explanations as features: Llm-based features for text-attributed graphs", in "In Proceedings of ICLR", (2024a).

He, X., Z. Lin, Y. Gong, H. Zhang, C. Lin, J. Jiao, S. M. Yiu, N. Duan, W. Chen *et al.*, "Annollm: Making large language models to be better crowdsourced annotators", North American Chapter of the Association for Computational Linguistics (NAACL) (2024b).

Ho, N., L. Schmid and S. Yun, "Large language models are reasoning teachers", in "61st Annual Meeting of the Association for Computational Linguistics, ACL 2023", pp. 14852–14882 (Association for Computational Linguistics (ACL), 2023).

Jain, S., S. Wiegreffe, Y. Pinter and B. C. Wallace, "Learning to faithfully rationalize by construction", in "58th Annual Meeting of the Association for Computational Linguistics, ACL 2020", pp. 4459–4473 (Association for Computational Linguistics (ACL), 2020).

Jeong, U., A. Nirmal, K. Jha, S. X. Tang, H. R. Bernard and H. Liu, "User migration across multiple social media platforms", in "Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)", pp. 436–444 (SIAM, 2024).

Jiang, B., Z. Tan, A. Nirmal and H. Liu, "Disinformation detection: An evolving challenge in the age of llms", in "Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)", pp. 427–435 (SIAM, 2024).

Kennedy, C. J., G. Bacon, A. Sahn and C. von Vacano, "Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application", arXiv preprint arXiv:2009.10277 (2020).

Kim, J., B. Lee and K.-A. Sohn, "Why is it hate speech? masked rationale prediction for explainable hate speech detection", in "Proceedings of the 29th International Conference on Computational Linguistics", pp. 6644–6655 (2022a).

Kim, Y., S. Park and Y.-S. Han, "Generalizable implicit hate speech detection using contrastive learning", in "Proceedings of the 29th International Conference on Computational LinguisticsProceedings of the 29th International Conference on Computational Linguistics", pp. 6667–6679 (2022b).

Kiritchenko, S., I. Nejadgholi and K. C. Fraser, "Confronting abusive language online: A survey from the ethical and human rights perspective", Journal of Artificial Intelligence Research **71**, 431–478 (2021).

Kumarage, T., A. Bhattacharjee and J. Garland, "Harnessing artificial intelligence to combat online hate: Exploring the challenges and opportunities of large language models in hate speech detection", arXiv preprint arXiv:2403.08035 (2024).

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations", in "International Conference on Learning Representations", (2019).

Latorre, J. P. and J. J. Amores, "Topic modelling of racist and xenophobic youtube comments. analyzing hate speech against migrants and refugees spread through youtube in spanish", in "Ninth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'21)", pp. 456–460 (2021).

Laub, Z., "Hate speech on social media: Global comparisons", Council on foreign relations **7** (2019).

Li, L., L. Fan, S. Atreja and L. Hemphill, ""hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media", ACM Transactions on the Web (2023).

Liu, H., P. Burnap, W. Alorainy and M. L. Williams, "Fuzzy multi-task learning for hate speech type identification", in "The world wide web conference", pp. 3006–3012 (2019a).

Liu, Y., 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", (2019b).

Lundberg, S. M. and S.-I. Lee, "A unified approach to interpreting model predictions", Advances in neural information processing systems **30** (2017).

Markov, I., N. Ljubešić, D. Fišer and W. Daelemans, "Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection", in "Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis", pp. 149–159 (2021).

Mathew, B., 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", vol. 35, pp. 14867–14875 (2021).

Miller, T., "Explanation in artificial intelligence: Insights from the social sciences", Artificial Intelligence **267**, 1–38 (2019).

Min, B., H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz and D. Roth, "Recent advances in natural language processing via large pre-trained language models: A survey", ACM Computing Surveys **56**, 2, 1–40 (2023).

Nandhini, B. S. and J. Sheeba, "Cyberbullying detection and classification using information retrieval algorithm", in "Proceedings of the 2015 international conference on advanced research in computer science engineering & technology (ICARCSET 2015)", pp. 1–5 (2015).

Nirmal, A., A. Bhattacharjee, P. Sheth and H. Liu, "Towards interpretable hate speech detection using large language model-extracted rationales", in "Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024), NAACL", (2024).

Nirmal, A., B. Jiang and H. Liu, "Sociohub: An interactive tool for cross-platform social media data collection", SBP-BRiMS (2023).

Nobata, C., J. Tetreault, A. Thomas, Y. Mehdad and Y. Chang, "Abusive language detection in online user content", in "Proceedings of the 25th international conference on world wide web", pp. 145–153 (2016).

Nockleby, J. T., "Hate speech in context: The case of verbal threats", Buff. L. Rev. **42**, 653 (1994).

Ocampo, N. B., E. Sviridova, E. Cabrio and S. Villata, "An in-depth analysis of implicit and subtle hate speech messages", in "Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics", pp. 1997–2013 (Association for Computational Linguistics, 2023).

Pan, L., C.-W. Hang, A. Sil and S. Potdar, "Improved text classification via contrastive adversarial training", in "Proceedings of the AAAI Conference on Artificial Intelligence", vol. 36, pp. 11130–11138 (2022).

Perera, S., N. Meedin, M. Caldera, I. Perera and S. Ahangama, "A comparative study of the characteristics of hate speech propagators and their behaviours over twitter social media platform", Heliyon **9**, 8 (2023).

Petroni, F., T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu and A. Miller, "Language models as knowledge bases?", in "Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)", pp. 2463–2473 (2019).

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners", OpenAI blog **1**, 8, 9 (2019).

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer", Journal of machine learning research **21**, 140, 1–67 (2020).

Ribeiro, M. T., S. Singh and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier", in "Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining", pp. 1135–1144 (2016).

Rodriguez, A., C. Argueta and Y.-L. Chen, "Automatic detection of hate speech on facebook using sentiment and emotion analysis", in "2019 international conference on artificial intelligence in information and communication (ICAIIC)", pp. 169–174 (IEEE, 2019).

Salminen, J., H. Almerekhi, M. Milenković, S.-g. Jung, J. An, H. Kwak and B. Jansen, "Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media", in "Proceedings of the International AAAI Conference on Web and Social Media", vol. 12 (2018).

Sap, M., D. Card, S. Gabriel, Y. Choi and N. A. Smith, "The risk of racial bias in hate speech detection", in "Proceedings of the 57th annual meeting of the association for computational linguistics", pp. 1668–1678 (2019).

Schmidt, A. and M. Wiegand, "A survey on hate speech detection using natural language processing", in "Proceedings of the fifth international workshop on natural language processing for social media", pp. 1–10 (2017).

Serra, J., I. Leontiadis, D. Spathis, G. Stringhini, J. Blackburn and A. Vakali, "Class-based prediction errors to detect hate speech with out-of-vocabulary words", in "Proceedings of the first workshop on abusive language online", pp. 36–40 (2017).

Sheth, P., T. Kumarage, R. Moraffah, A. Chadha and H. Liu, "Peace: Cross-platform hate speech detection-a causality-guided framework", in "Joint European Conference on Machine Learning and Knowledge Discovery in Databases", pp. 559–575 (Springer, 2023).

Sheth, P., R. Moraffah, T. S. Kumarage, A. Chadha and H. Liu, "Causality guided disentanglement for cross-platform hate speech detection", in "Proceedings of the 17th ACM International Conference on Web Search and Data Mining", pp. 626–635 (2024).

Singh, C., J. P. Inala, M. Galley, R. Caruana and J. Gao, "Rethinking interpretability in the era of large language models", arXiv preprint arXiv:2402.01761 (2024).

Tarunesh, I., S. Aditya and M. Choudhury, "Trusting roberta over bert: Insights from checklisting the natural language inference task", ArXiv **abs/2107.07229**, URL `https://api.semanticscholar.org/CorpusID:235899209` (2021).

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models", (2023).

Ullmann, S. and M. Tomalin, "Quarantining online hate speech: technical and ethical perspectives", Ethics and Information Technology **22**, 1, 69–80 (2020).

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

Warner, W. and J. Hirschberg, "Detecting hate speech on the world wide web", in "Proceedings of the second workshop on language in social media", pp. 19–26 (2012).

Weir, G., K. Owoeye, A. Oberacker and H. Alshahrani, "Cloud-based textual analysis as a basis for document classification", in "2018 International Conference on High Performance Computing & Simulation (HPCS)", pp. 672–676 (IEEE, 2018).

Wiedemann, G., S. M. Yimam and C. Biemann, "Uhh-lt at semeval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection", in "Proceedings of the Fourteenth Workshop on Semantic Evaluation", pp. 1638–1644 (2020).

Yang, J., C. Liu, W. Deng, D. Wu, C. Weng, Y. Zhou and K. Wang, "Enhancing phenotype recognition in clinical notes using large language models: Phenobcbert and phenogpt", Patterns **5**, 1 (2024).

Yin, W., V. Agarwal, A. Jiang, A. Zubiaga and N. Sastry, "Annobert: Effectively representing multiple annotators' label choices to improve hate speech detection", in "Proceedings of the International AAAI Conference on Web and Social Media", vol. 17, pp. 902–913 (2023).

Zannettou, S., M. Sirivianos, J. Blackburn and N. Kourtellis, "The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans", Journal of Data and Information Quality (JDIQ) **11**, 3, 1–37 (2019).

Zhu, Y., H. Yuan, S. Wang, J. Liu, W. Liu, C. Deng, Z. Dou and J.-R. Wen, "Large language models for information retrieval: A survey", ArXiv **abs/2308.07107**, URL `https://api.semanticscholar.org/CorpusID:260887838` (2023a).

Zhu, Y., P. Zhang, E.-U. Haq, P. Hui and G. Tyson, "Can chatgpt reproduce human-generated labels? a study of social computing tasks", arXiv e-prints pp. arXiv–2304 (2023b).

APPENDIX A

CSS EXPLORATORY RESEARCH IN THE FIELD OF LLMS

1. Disinformation Detection in the Era of Large Language Models

The emergence of Large Language Models (LLMs), such as ChatGPT (Brown *et al.*, 2020) and LLaMA (Touvron *et al.*, 2023), marks a pivotal advancement in the field of Computational Social Science (CSS). These models have revolutionized our ability to analyze human language and behavior. However, with this progress comes a pressing concern: the potential for these models to be exploited for disinformation generation and dissemination. As LLMs continue to evolve, reaching levels of generating content that is indistinguishable from human-produced text, the specter of AI-generated disinformation looms large. Indeed, recent studies have underscored the alarming efficiency and effectiveness of such disinformation campaigns. In the pre-LLM era, research in AI-generated disinformation detection primarily centered around Smaller Language Models (SLMs) like BERT (Vaswani *et al.*, 2017), GPT-2 (Radford *et al.*, 2019), and T5 (Raffel *et al.*, 2020). However, the advent of LLMs, with their billion-scale parameters, has drastically increased the complexity of disinformation detection. Textual outputs from LLMs exhibit naturalness and human-like qualities, posing significant challenges to existing detection techniques designed around SLMs. Despite the importance of this shift, its ramifications remain largely unexplored.

In our paper (Jiang *et al.*, 2024), we aim to address this gap in understanding by investigating the applicability of existing disinformation detection techniques to LLM-generated content. We pose the following research questions:

1. Are current disinformation detection methods suitable for identifying LLM-generated disinformation?

2. If not, can LLMs themselves be repurposed to detect such disinformation?

3. If traditional methods and LLM-based approaches fall short, what alternative strategies can be considered?

To ground our investigation in practical relevance, we contextualize our research within a hypothetical scenario wherein malicious actors leverage LLMs to produce sophisticated disinformation campaigns aimed at subverting automated detection systems and influencing public opinion. We utilize benchmark datasets of human-written news articles categorized as true or fake, constructing novel disinformation datasets of varying complexity levels using ChatGPT (GPT-3.5 and 4) and three prompt techniques.

Our approach involves evaluating the efficacy of state-of-the-art disinformation detection methods, initially fine-tuning a RoBERTa-based model on human-written disinformation datasets. Subsequently, we assess its performance in detecting LLM-generated disinformation. We also explore the inherent capabilities of LLMs in discerning self-generated disinformation and propose innovative methods inspired by human fact-checking processes.

Through extensive experimentation, we make several key observations:

1. Existing detection methods struggle to identify higher complexity LLM-generated disinformation.

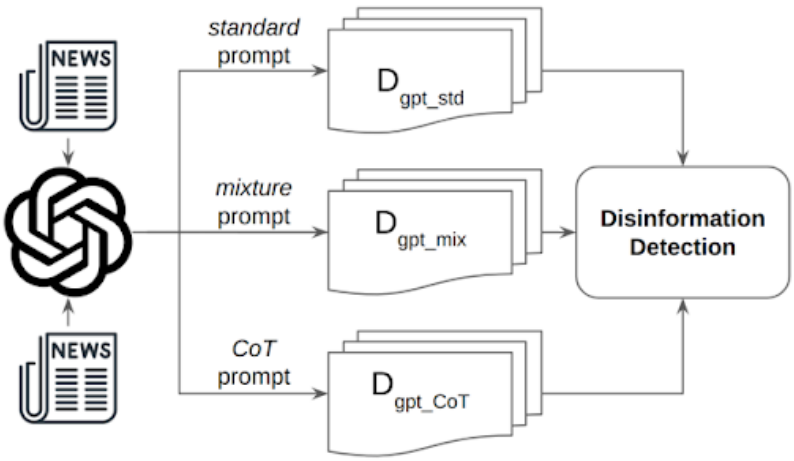

Figure A.1: Overview of Disinformation Generation and Detection Using Chatgpt. We Input Human-crafted Disinformation (Left) along with Distinct Prompts to Produce Three Separate Llm-generated Disinformation Datasets (Center). We Subsequently Evaluate the Efficacy of the Disinformation Detection System (Right) Against Llm-generated Disinformation (Jiang *et al.*, 2024).

2. Vanilla ChatGPT demonstrates limited efficacy in detecting even its own generated disinformation.

3. However, leveraging carefully designed prompts, inspired by chain-of-thought techniques, shows promising improvements in detection accuracy.

Our work contributes to the field through dataset curation, problem validation highlighting the inadequacy of current detection techniques, and the proposal of novel frameworks for LLM-generated disinformation detection.

2. LLMs as Pseudo Stance Detection Annotators

In addressing the challenge of assessing users' brand loyalty towards various social media platforms, particularly amidst shifts in ownership and policy changes, we employed Large Language Models (LLMs) as pseudo annotators for stance detection (Jeong *et al.*, 2024). Given the absence of annotated datasets tailored specifically for brand loyalty detection, we utilized ChatGPT (GPT-4) due to its established proficiency in stance detection tasks. With a custom prompt tailored for target-based stance detection, we analyzed users' sentiments towards the studied platforms, categorizing stances into loyalty, neutrality, or disloyalty. This process involved extracting posts mentioning specific platforms, chronologically grouping them by user, and subsequently annotating the stances of a sample of 400 users. Notably, the agreement between LLM annotations yielded a significant Cohen's Kappa coefficient of 76.27%, indicating a high level of agreement with human annotators. We further validated this approach by achieving an F1 score of 79.42%, demonstrating the effectiveness of utilizing LLMs for stance detection in this context. Through structured inputs

and outputs in a few-shot setting, leveraging the capabilities of GPT-4 via OpenAI API's function calling feature, we incorporated pertinent information about platform rebranding and new social media launches, ensuring comprehensive coverage in assessing users' sentiments towards these platforms amidst shifting landscapes.

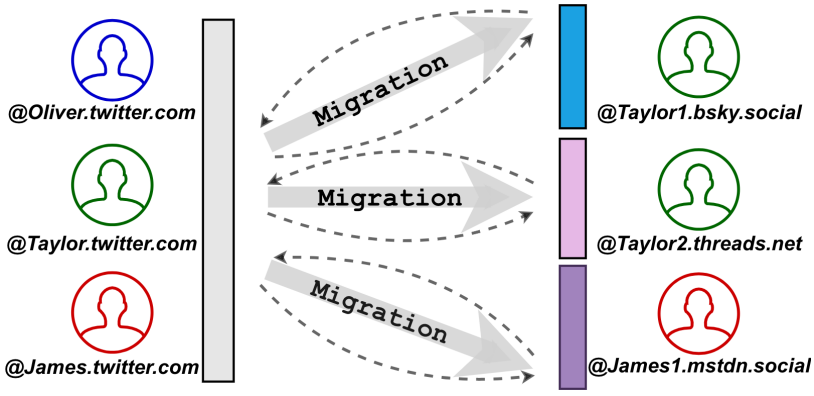

Figure A.2: The Migration Flow Between Twitter and Its Alternatives: Mastodon, Bluesky, and Threads. The Dashed Lines Represent the Shift of User Attention Across These Platforms (Jeong *et al.*, 2024).

Title: Stance Detection Prompt for GPT-4
Objective:
Determine the stance of a given text towards a specified target platform.
Instructions:
For the provided text and target(s), classify the stance as one of the following: Loyal, Disloyal, or Neutral
Keynotes:
- There are four platforms that can be targeted for the stance: Twitter, Bluesky, Threads, and Mastodon.
- Twitter is now called "𝕏" and is owned by Elon Musk.
- "Threads" is a new social media platform under the Meta umbrella, founded by Mark Zuckerberg.
- Always provide a stance for each specified target.
- Provide the response in JSON format.
Example:
Input: "Twitter is dead. I love Bluesky"
Output: {Twitter: Disloyal, Bluesky: Loyal}