universität wien

# MASTERARBEIT | MASTER'S THESIS

Titel | Title

## Hybrid Approaches in Implicit Hate Speech Detection: GPT-Driven Feature Extraction and Supervised Learning

verfasst von | submitted by

Julia Meta Pardatscher BA

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2024

| | |
|---|---|
| Studienkennzahl lt. Studienblatt \| Degree programme code as it appears on the student record sheet: | UA 066 587 |
| Studienrichtung lt. Studienblatt \| Degree programme as it appears on the student record sheet: | Joint-Masterstudium Multilingual Technologies |
| Betreut von \| Supervisor: | Assoz. Prof. Mag. Dr. Dagmar Gromann BSc |

# Acknowledgements

I would like to express my deepest gratitude to all those who supported me throughout my journey to complete this master's thesis.

Firstly, I am profoundly grateful to Dr. Michael Wiegand for his invaluable guidance, insightful feedback, and continuous encouragement. His expertise and commitment played an instrumental role in shaping this thesis.

Secondly, I express my sincere gratitude to my supervisor, Assoc. Prof. Dr. Dagmar Gromann, for her thoughtful suggestions, mentorship, and support. My peers and I are well aware of the tremendous effort Prof. Gromann has put into the Multilingual Technologies program.

I would also like to thank my family and friends for their understanding and belief in me. Their patience and trust provided the strength I needed to persevere through challenging times.

A special thank you to my partner, Michael Jaremko, for his constant encouragement and understanding. His unwavering support was vital to my success.

I am grateful to my colleagues and peers in the Multilingual Technologies program. Their wonderful humor, collaborative spirit, and meaningful discussions enriched my research experience.

I would also like to thank the Digital Age Research Center at the University of Klagenfurt for providing the data and resources necessary for this study. Without their cooperation, this research would not have been possible.

I extend my sincere thanks to the Vienna Employment Promotion Fund for awarding me a generous scholarship, which provided significant financial support and allowed me to focus on my research.

Finally, I am grateful to the University of Vienna and the University of Applied Sciences Campus Vienna for providing a stimulating academic environment.

Thank you all for your support and encouragement.

# Abstract

Hate speech, particularly in its implicit form, is a pervasive issue in online communication, necessitating sophisticated detection methods to foster a respectful digital environment. Unlike explicit hate speech, which is overtly offensive, implicit hate speech is characterized by covert and ambiguous expressions that are often ironic or euphemistic. Linguistic features that characterize such forms of hate speech include the use of extreme imagery, taboo topics, or contradictory comparisons that convey a hostile attitude without using overtly offensive language.

Despite extensive research, detecting implicit hate speech remains a major challenge because these subtle linguistic nuances are difficult to discern. Traditional approaches to automatically detecting implicit hate speech often rely on manually annotated datasets where linguistic features were labeled by human annotators. Such approaches depend on accurately annotated training data. However, this dependency is problematic because the manual annotation process is time-consuming and costly, especially for smaller companies and organizations.

This thesis proposes an alternative approach by replacing human annotators with Generative Pre-Trained Transformer (GPT) models, thereby facilitating the extraction of linguistic features. The goal is to improve the performance of implicit hate speech classifiers while minimizing dependence on hand-labeled data. The experiments focus on hate speech in English and specifically use the LLaMA-2, GPT-3.5, and GPT-4 models.

The GPT models were prompted to annotate multiple datasets using zero-shot and few-shot prompting methods. Specifically, the GPT models were directed to extract the linguistic features of implicit hate speech from four public datasets. This automatically annotated data was then used to train supervised machine learning models. To evaluate the efficiency of this method, the same models were trained on data annotated exclusively by humans. By comparing the results, it can be determined whether the data extracted by GPT can match or even exceed the quality of the manually annotated data.

This proposed hybrid approach combines GPT-driven feature extraction with supervised feature-based machine learning methods, simplifying the often complex process of feature extraction.

The results of the study showed clear differences in the efficiency of the different GPT models. The performance of LLaMA-2 was unsatisfactory. In comparison, GPT-3.5 achieved better results but did not surpass the quality of the manually annotated data. The results of the classifiers trained with features extracted by GPT-4 show an improvement in detection capabilities compared to classifiers trained on human annotations.

Ultimately, the results indicate that GPT models can be a promising alternative to manual annotation, especially in the detection of implicit hate speech. The use of such models can significantly lessen the resource burden of manual annotation. This reduces the need for extensive manual work, thereby minimizing cost and effort while improving recognition accuracy.

**Given the sensitive nature of this research, it is important to acknowledge the challenges of working with hostile language. The material analyzed includes**

**disturbing and toxic content, and the examples provided may be troubling and potentially offensive to readers. Nevertheless, openly discussing real-world instances of hate speech is crucial for understanding its mechanisms and ultimately finding solutions to address it.**

# Zusammenfassung

Hassrede, insbesondere in ihrer impliziten Form, stellt eine ernsthafte Bedrohung dar und kann weitreichende negative Auswirkungen auf Einzelpersonen und Gemeinschaften haben. Im Gegensatz zu expliziter Hassrede, die offen beleidigend ist, zeichnet sich implizite Hassrede durch versteckte und mehrdeutige Ausdrücke aus, die oft ironisch oder euphemistisch sind. Linguistische Merkmale, die solche Formen der Hassrede kennzeichnen, umfassen z.B. die Verwendung extremer bildlicher Sprache, tabuisierter Themen oder widersprüchlicher Vergleiche, die eine feindselige Haltung vermitteln, ohne offensichtlich beleidigende Ausdrücke zu nutzen.

Trotz umfangreicher Forschung bleibt die Erkennung impliziter Hassrede eine große Herausforderung, da diese subtilen linguistischen Nuancen schwer zu identifizieren sind. Herkömmliche Ansätze zur automatisierten Erkennung impliziter Hassrede stützen sich häufig auf manuell annotierte Datensätze, in denen linguistische Merkmale von menschlichen Annotator*innen gekennzeichnet wurden. Solche Ansätze sind auf ausreichend genau annotierte Trainingsdaten angewiesen. Diese Abhängigkeit stellt allerdings ein Problem dar, da der manuelle Annotationsprozess zeitaufwändig und kostspielig ist, insbesondere für kleinere Unternehmen und Organisationen.

Diese Masterarbeit schlägt vor, menschliche Annotator*innen durch generative vortrainierte Transformermodelle (GPT) zu ersetzen, um die Effizienz und Genauigkeit von automatisierten Erkennungsmethoden zu verbessern. Ziel ist es, die Abhängigkeit von manuell annotierten Daten zu reduzieren und gleichzeitig die Erkennungsgenauigkeit zu erhöhen. Die Experimente konzentrieren sich auf Hassrede im Englischen und verwenden speziell die Modelle LLaMA-2, GPT-3.5 und GPT-4.

Die GPT-Modelle wurden durch Zero-Shot- und Few-Shot-Prompting angewiesen, mehrere Datensätze zu annotieren. Genauer gesagt, wurden die GPT-Modelle dazu angewiesen, die linguistischen Merkmale impliziter Hassrede aus drei öffentlichen Datensätzen zu extrahieren. Diese automatisch annotierten Daten wurden dann zum Trainieren maschineller Lernmodelle verwendet. Um die Effizienz dieser Methode zu evaluieren, wurden dieselben Modelle mit Daten trainiert, die ausschließlich von Menschen annotiert wurden. Der Vergleich der Ergebnisse erlaubt es festzustellen, ob die von GPT extrahierten Daten die Qualität der manuell annotierten Daten erreichen oder sogar übertreffen können.

Die Ergebnisse der Untersuchung zeigten deutliche Unterschiede in der Effizienz der verschiedenen Modelle. Die Leistung von LLaMA-2 war nicht zufriedenstellend. Im Vergleich dazu erzielte GPT-3.5 bessere Ergebnisse, die jedoch die Qualität der menschlich annotierten Daten nicht übertrafen. Hervorzuheben sind die Resultate von GPT-4, dessen generierte Annotationen zu einer signifikanten Verbesserung der Trainingsmodelle führten.

Schließlich zeigen die Ergebnisse, dass GPT-Modelle eine vielversprechende Alternative zur manuellen Annotation darstellen können, insbesondere in der Erkennung impliziter Hassrede. Der Einsatz solcher Modelle kann die Ressourcenbelastung durch manuelle Annotationen erheblich verringern. Dies reduziert die Notwendigkeit für umfangreiche manuelle Arbeiten, was Kosten und Aufwand minimiert und gleichzeitig die Erkennungsgenauigkeit verbessert.

# Contents

# List of Tables

# List of Figures

# 1. Introduction

Hate speech, characterized by its targeting of individuals or groups based on protected characteristics such as sexual orientation, ethnicity, gender, or religion, inflicts not only emotional and physical harm but also perpetuates derogatory stereotypes (Brown, 2017a). Consequently, hate speech fosters an environment of discrimination, hostility, and societal division (Delgado, 1982; Matsuda, 1989). This is particularly concerning because the advent of computer-mediated communication has greatly increased the visibility and spread of discourse that was previously confined to private or marginal public spaces (Knoblock, 2022). Indeed, the evolving landscape of online communication presents a significant challenge in identifying nuanced forms of hate speech.

Whether hate speech should be banned or not is beyond the scope of this thesis. Nevertheless, the application of Natural Language Processing (NLP) methods to monitor hate speech presents a valuable tool for understanding and addressing its impact within democratic frameworks. After all, hate speech has been identified as a precursor to various violent acts and adverse mental health outcomes. Mullen and Smyth's (2004) study found a significant correlation between the suicide rates among ethnic immigrant groups in the U.S. and the negativity of the racial slurs directed at them. Specifically, the research indicated that groups subjected to more negative and complex forms of hate speech exhibited higher suicide rates. Similarly, Müller and Schwarz's (2017) findings suggest that social media can aid in the spread of extreme views, which in turn can lead to violent actions. They demonstrate that increased anti-refugee sentiment expressed on Facebook (now Meta) is correlated with a higher incidence of crimes against refugees in municipalities that exhibit higher levels of social media usage, in contrast to similar municipalities with lower social media usage. Automated tools could provide a scalable and efficient means to monitor and manage the vast amounts of online communication, potentially identifying and mitigating hate speech before it escalates into real-world violence and adverse mental health outcomes.

However, current research predominantly focuses on explicit hate speech, yet the subtler manifestations of implicit hate speech remain a substantial concern (van Aken et al., 2018; Wiegand et al., 2019, 2021b). Implicitly offensive language is characterized by its veiled nature, requiring a more sophisticated approach for identification. In this thesis, explicit and implicit hate speech are distinguished in the following manner. Explicit instances, such as those shown in Example (1), involve overtly offensive language targeting a specific group. In contrast, implicit cases, like those in Example (2), entail subtler expressions that convey hostility without using explicit discriminatory language (Wiegand et al., 2022, 5600-5601).

(1) "Go lick a pig, you Arab Muslim piece of scum."

(2) "Jews succumb to cultural degeneracy."

Automated detection of implicit hate speech is more challenging than identifying explicit instances because it requires classifiers to go beyond simple keyword detection and recognize complex, subtle linguistic patterns. This challenge highlights a broader issue in supervised learning, which is the dependence on well-annotated data for effective model training (Plaza-del arco et al., 2023). Supervised learning relies on having enough accurately annotated training data, and the annotation process can be costly, especially for smaller companies and organizations, as indicated by Ding et al. (2022). These costs include labor for data tagging, hiring and training annotators, and expenses related to annotation tools and infrastructure. Smaller entities may lack the resources to produce sufficient training data, hindering their ability to use advanced modeling techniques. Although pre-trained language models like Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), XLNet (Yang et al., 2019), and Robustly Optimized BERT Pre-training Approach (RoBERTa) (Liu et al., 2019) alleviate some of the data requirements, Ding et al. (2022) stress that data annotation remains a crucial and inevitable challenge for supervised model training. Generative Pre-trained Transformers (GPT) have the potential to be a promising alternative to human annotators. They can generate text that closely mimics human language and perform various NLP tasks, such as translating languages, creating summaries, and answering questions (Ding et al., 2022).

This leads to the central research question of this master's thesis. How effectively can GPT models extract linguistic features of implicit hate speech in English? How can those predictions contribute to the automated detection of implicit hate speech?

First, the effectiveness of the GPT models LLaMA-2 (Touvron et al., 2023), GPT-3.5 (Brown et al., 2020a), and GPT-4 (OpenAI et al., 2023) in directly recognizing implicit hate speech was evaluated. For this purpose, four datasets were selected: *Identity Groups* (Wiegand et al., 2022), *Euphemistic Abuse* (Wiegand et al., 2023), *Comparisons* (Wiegand et al., 2021a), and *ISHate* (Ocampo et al., 2023). Using both zero-shot and few-shot prompting approaches, the GPT models were instructed to classify sentences as either implicitly abusive language or benign utterances. Following this, the focus shifted to the automated extraction of linguistic features of implicit hate speech using GPT models. Despite its subtle nature, implicit hate speech has identifiable characteristics that render it amenable to automated detection. Typically, these characteristics or linguistic features are manually extracted by human annotators (Mollas et al., 2020; Founta et al., 2018). The aim of this thesis was to test if GPT models could extract these linguistic features as well. To evaluate the quality of the extracted features, macro-averaged F1 scores were analyzed. Additionally, logistic regression models were trained on these automatically extracted features and compared with models trained on manually extracted features. This process was also applied to a basic rule-based classifier. Results from the GPT-generated features were compared with those from a rule-based classifier using manually generated features.

The findings indicate that LLaMA-2 exhibited poor performance in detecting implicit hate speech. In comparison, GPT-3.5 achieved better results but still struggled with recognizing implicitly abusive language and did not improve with few-shot prompting. Conversely, GPT-4 demonstrated strong capabilities in zero-shot detection and showed further improvement with few-shot learning. The logistic regression models showed improved performance when incorporating linguistic features extracted by GPT-4, compared to identical models trained on manually extracted features, as indicated by the evaluation scores. Similarly, the rule-based classifier demonstrated performance improvements when using features that were extracted automatically by GPT models, rather than relying on manually extracted features. These findings suggest that GPT-4, in particular, offers a viable alternative to human annotators.

Following this introduction, Chapter 2 examines the definitions, impacts, and debates surrounding hate speech, providing a contextual foundation. Chapter 3 discusses the technical scope, including neural language models and existing methods for hate speech detection. Chapter 4 reviews related work, highlighting advancements in Large Language Models (LLM) for hate speech detection. Chapter 5 details the method, focusing on data collection, model training, and feature extraction using GPT models. Chapter 6 presents the results of the experiments, showcasing the performance of the proposed hybrid models. Chapter 7 offers a discussion on the findings, limitations, and ethical considerations. Finally, Chapter 8 concludes the thesis, summarizing the key contributions and suggesting directions for future research. The code for the presented experiments is available at https://bit.ly/implicit-hate-speech-detection-and-feature-extraction-using-gpt-models.

# 2. Understanding and Addressing Hate Speech: Definitions, Impacts, and Debates

The first chapter is organized into three parts. It begins by clarifying what hate speech is, laying the groundwork for identifying it. The second section discusses the harms caused by hate speech and its impact on individuals and society. Finally, the third section navigates the complex debate between prohibiting hate speech and preserving freedom of expression.

## 2.1. Defining Hate Speech from Legal, Linguistic, and Social Perspectives

This section attempts to define hate speech by examining it from multiple perspectives. It begins with an exploration of legal definitions and interpretations. Following this, it investigates how hate speech is addressed within NLP, including definitions proposed by researchers and guidelines used by social media platforms. Finally, the linguistic and social dimensions of hate speech are considered, analyzing its usage, implications, and the difficulties in establishing a universally accepted definition.

The term hate speech itself was conceived in the late 1980s by a team of legal scholars in the United States, reacting to the various legal approaches to managing certain types of harmful racist expressions (Matsuda, 1989). This term has since transcended its original legalistic confines, entering the lexicon of both the media and the general public, and evolving significantly in its application and interpretation. Traditionally, hate speech has been a term predominantly used by liberal progressives or politically left-leaning stakeholders to identify and disparage expressions deemed racist, xenophobic, homophobic, Islamophobic, misogynistic, disablist, or otherwise discriminatory towards identity groups, arguing that such speech violates principles of respect, solidarity, tolerance, etc. (Brown, 2017a). In this context, identity groups are understood as collections of individuals who see themselves as part of the same social category, share an emotional bond to this common identity, and agree on the significance of their group and their belonging to it (Tajfel and Turner, 1979).

Within the legal landscape, the concept of hate speech encompasses a variety of interpretations and definitions. Notably, in cases such as Surek v. Turkey (European Court of Human Rights, 1999) and Gündüz v. Turkey (European Court of Human Rights, 2003), the European Court of Human Rights (ECtHR) applied a definition originally established by the Committee of Ministers of the Council of Europe in 1997 (Council of Europe, Committee of Ministers, 1997). This definition encompasses all forms of expression that spread, incite, promote, or justify racial hatred, xenophobia, antisemitism, or other forms of intolerance-based hatred. In both aforementioned instances, the ECtHR highlighted that for expressions to be classified as hate speech, they must go beyond merely disturbing, offending, or shocking; they must have the capacity to incite violence or hatred.

Similarly, domestic legislation in South Africa outlines hate speech offenses:

> [...] no person may publish, propagate, advocate or communicate words based on one or more of the prohibited grounds, against any person, that could reasonably be construed to demonstrate a clear intention to: (a) be hurtful; (b) be harmful or to incite harm; (c) promote or propagate hatred. (Republic of South Africa, 2000, 9)

Additionally, the South African Bill of Rights, under s 16(2), states that freedom of expression does not extend to the promotion of hatred based on race, ethnicity, gender, or religion that incites harm (Constitutional Assembly of South Africa, 1996).

Beyond these examples, the legal definitions of hate speech are vast and varied. They often extend into analyses of legal concepts through terminologies not explicitly labeled as hate speech, but are closely related or serve as proxies. Such terms include hate, hatred, contempt, hostility, enmity, feelings of inferiority, racist propaganda, xenophobia, anti-semitism, aggressive nationalism, homophobia, Islamophobia, group defamation, group vilification, insult, negative stereotyping, stigmatization, humiliation, degradation, dignity violations, discriminatory harassment, and intolerance. These diverse legal interpretations highlight the complex nature of identifying and regulating hate speech within various legal frameworks (Brown, 2017a).

Notably, the term hate speech is the most recent in a series of terms historically used to describe expressions that target individuals or groups based on certain protected characteristics, such as race hate, group libel, and hate propaganda. These terms have been instrumental for societies in addressing and categorizing forms of prejudicial speech. However, what sets hate speech apart is its broad scope. It is capable of encompassing a wider array of protected characteristics and forms of speech than its predecessors. Malleson (2018) define protected characteristics as attributes recognized by legislation, safeguarding people from discrimination or unjust treatment due to those attributes. For a characteristic to be protected, it must meet three key criteria: it should have some definitional and categorical stability, reflect a broad understanding of social realities and lived experiences, and correspond with the most significant axes of discrimination present in society (Malleson, 2018). This broad applicability of hate speech is critical, as it refers to a more expansive concept, capturing the expressive dimensions of identity-driven mistrust, conflict, envy, animosity, and oppression (Brown, 2017a).

From an NLP perspective, Fortuna and Nunes (2018) gathered and analyzed definitions of hate speech from various sources, including the European Union Commission, which sets guidelines for its institutions; International minorities associations like the International Lesbian, Gay, Bisexual, Trans and Intersex Association (ILGA), which focus on protecting groups often targeted by hate speech; and the scientific community, offering a research-based viewpoint. Additionally, they looked at the conditions and terms of social networks, such as Facebook, YouTube, and Twitter (now called X), platforms where hate speech frequently occurs. This comprehensive analysis led Fortuna and Nunes (2018) to propose their definition of hate speech.

> Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity

> or other, and it can occur with different linguistic styles, even in subtle forms or when humour is used. (Fortuna and Nunes, 2018, 85:5)

To aid in the identification of hate speech, Fortuna and Nunes (2018) further propose a set of main rules to identify hate speech. Hate speech is present when individuals:

- highlight an individual's group membership and attach a disparaging stereotype to that group,
- make generalized negative remarks about minority groups, stirring a negative bias towards them,
- utilize disparaging terms and racial epithets intending to cause harm,
- employ sexist or racial slurs,
- use sexist or racial slurs to express pride, even if the speaker belongs to the group targeted by these slurs. However, this classification can change if the context clearly shows the speaker's group membership. If there is no clear contextual indication of the speaker's group membership, such terms are categorized as hateful.
- Support organizations known for advocating hate speech, even without making direct verbal attacks, or
- assert the superiority of one's own group over others.
- Criticizing countries or religions is generally allowed, but discrimination based on these aspects is not.

Brown (2017a) offers a series of valuable definitions from the linguistic perspective. The approach begins by treating hate speech as a complex or compositional concept, which is built from more basic, simpler concepts. These foundational concepts cover broader categories, allowing complex concepts to be broken down through decompositional conceptual analysis. This method involves dissecting the complex concept into its basic components to establish a precise definition. For instance, the concept of a bachelor is clarified by combining three simpler concepts: being an adult human, being unmarried, and being male.

Subsequently, understanding the concept of hate speech as a compositional concept raises the question of how to identify its component parts. For example, one might attempt an Aristotelian definition. Such a definition, named after the ancient Greek philosopher Aristotle, seeks to explain the essence of an entity by specifying its genus, a general category or class to which the entity belongs, and its differentia, the specific characteristic that distinguishes it from other members of its genus. This form of definition is also known as a genus-differentia definition. It aims to provide a concise explanation of what something is by highlighting its fundamental nature within a broader category and identifying what makes it unique within that category (Roelcke, 2010).

One could for example argue that the term hate speech functions as a hyponym of speech, much like olive oil is a specific type within the broader category of oil (Brown, 2017a). A hyponym is a word that represents a subcategory of a more general class also called hypernym, indicating a specific type of the broader category. In the context of genus-differentia definitions, the concept being defined can be seen as a hyponym of its genus. This suggests that hate speech represents a distinct subset or hyponym of the genus speech. Furthermore, the inclusion of hate within the term hate speech may serve a semantic purpose or differentia, signaling the essential nature of this subcategory of speech. Specifically, it indicates that hate speech is deeply intertwined with emotions, feelings, or attitudes of hate or hatred. Thus, the term hate speech not only identifies a particular kind of speech but also infuses it with a specific emotional context, highlighting its association with feelings of hate or hatred. Recognizing hate speech as a hyponym emphasizes its function in specifying a distinct segment of speech, defined by its content and underlying intention.

Yet, this interpretation of hate speech as merely a subset of speech linked to hateful emotions remains insufficiently precise for the requirements of this master's thesis. The term hate speech encapsulates complexities beyond the simple conjoining of hate and speech. Thus, a deeper exploration into its composition is necessary, as suggested by Brown's (2017a) decompositional conceptual analysis.

According to Brown (2017a), the concept of hate speech could instead integrate three foundational, simpler concepts: first, speech or other forms of expressive conduct; second, the identification of groups or classes of persons by protected characteristics; and third, the embodiment of emotions, feelings, or perspectives of hate or hatred. Importantly, the simultaneous presence of these conditions serves as a comprehensive criterion for labeling an act as hate speech.

So far, definitions from legal, NLP, and linguistic perspectives have been examined. These perspectives share a common foundation: they all rely on the "myth of hate" (Brown, 2017a, 432). This myth theorizes that for speech to be classified as hate speech it must be linked in some significant manner to emotions, feelings, or attitudes of hate or hatred. Such emotions are characterized by a profound or intense aversion, dislike, antipathy, loathing, hostility, or enmity toward something or someone, possibly accompanied by a desire to eliminate or banish the target of these emotions.

However, not all communications identified as hate speech necessarily originate from emotions, feelings, or attitudes of hate or hatred towards individuals or groups with protected characteristics. For instance, certain expressions of hate speech may stem from feelings of contempt, disdain, scorn, condescension, or dismissiveness. These feelings entail viewing someone or something as unworthy of consideration or respect, leading to a desire to withdraw from, avoid, or shun the target. In these scenarios, the speaker does not actually hate the subject of their speech; rather, they view them with contempt, considering them beneath the level of consideration or respect required to evoke feelings of hate or any significant regard (Brown, 2017a).

Consider the example provided by Brown (2017a) of a scientist who publishes findings on the comparative intelligence of African Americans and white Americans, stating that African Americans tend to have lower IQs than white Americans. This statement may not stem from any hate or hatred towards African Americans, whether consciously or unconsciously. It could represent a genuinely held belief based on the scientist's interpretation of data regarding genetic and environmental influences on interracial IQ differences and their implications for educational attainment, income, and social behaviors. This scenario raises a dilemma about whether such a statement qualifies as hate speech. On the one hand, the lack of hateful intent could be used as an argument against this categorization. On the other hand, the statement could be perceived as hate speech by many, including but not limited to African Americans, because it publicly reinforces rather than challenges a harmful stereotype or social stigma portraying African Americans as less intelligent than white Americans (Brown, 2017a).

Moreover, asserting that hate speech is uniformly motivated by hate oversimplifies the issue and overlooks the diverse and complex motivations behind such expressions. Brown (2017a) illustrates this point by providing several examples where speech that might instinctively be labeled as hate speech is not driven by hate or hatred. For instance, an individual might insult or mock members of the Muslim community not out of hate, but due to fear, a sense of loss, or feelings of alienation triggered by the presence of what they perceive as foreigners in their community. This suggests that actions commonly identified as hate speech can be motivated by a wide array of reasons that do not necessarily relate to hatred or even specifically target the group in question. According to Brown (2017a), the emotions experienced by the authors of hate speech can vary widely, as can their motivations, which include boredom, a desire for attention, enjoyment of controversy, and even reasons as pragmatic as economic self-interest. Consider another example provided by Brown (2017a), namely the case of a shopkeeper who spreads false and harmful statements about Jews in their area, not out of animosity towards Jews but because they view Jewish-owned businesses as competitors. Their aim is to divert the customers of the Jewish-owned businesses to their own business through a smear campaign.

Furthermore, the conceptualization of hate speech as merely the expression of hate or hatred can potentially backfire as it risks pathologizing hate speech (Brown, 2017a). Indeed, this perspective implies that hate speech is an aberration, a manifestation of mental pathology rather than a societal issue. By defining hate speech as the act of individuals overwhelmed by such intense hatred or deep-seated animosity towards certain groups that they are compelled to express this hatred vocally, it paints the phenomenon as abnormal. It characterizes the individuals who engage in hate speech as being driven by obsessional, paranoid, or irrational feelings, emotions, or attitudes, coupled with uncontrollable impulses to communicate these sentiments.

Yet, this perspective neglects the fact that many instances of hate speech are executed by individuals who cannot be deemed pathological in any clinical sense (Brown, 2017a). These individuals may not necessarily harbor hatred or deep-seated animosity towards others. Even if they do possess such feelings, they are often fully capable of controlling whether to express them. Hate speech, rather than being an anomaly occurring at the fringes of normal psychological behavior, is conducted by ordinary people making deliberate choices about their actions and words. Consequently, framing hate speech too narrowly as a product of hate or mental pathology fails to capture the broader, more complex social dynamics at play.

The final potential drawback of defining hate speech strictly as speech motivated by hate or hatred concerns the implications for legal regulation (Brown, 2017a). If hate speech laws are framed around this definition, they could be perceived as attempts to regulate individuals' emotions, feelings, or attitudes. This perspective raises concerns about the appropriateness of state intervention in the personal and private realms of thought and emotion. Many hold the instinctive belief that the state should not intrude into the inner world of its citizens, arguing that it is not within the state's purview to legislate emotions and feelings any more than it should legislate attitudes or thoughts, even those involving hate or hatred.

The discussion so far has aimed to point out the limitations of the assumption that the most accurate way to define hate speech is through a compositional analysis, particularly by relying on the literal meaning of hate as encompassing emotions, feelings, or attitudes. However, it could be helpful at this point to question the presumption that the term hate in hate speech retains its usual or literal meaning. This assumption might stem from observing the semantic consistency across various complex terms incorporating the word hate, such as hate tweets, hate campaign, hate propaganda, haters gonna hate, hate mail, hate crime, the politics of hate, etc. (Brown, 2017a). Yet, the critical issue here is that not all compound terms containing hate necessarily incorporate the ordinary or literal meaning of hate in their semantic composition. Therefore, it would be a mistake to assume that hate functions semantically in hate speech in the same manner as it does in other complex terms that also include the word.

Brown (2017a) ultimately proposes that the term hate speech might best be understood as a relational metaphor. This interpretation suggests that hate speech conveys a relational structure common to both the feelings or emotions of hate or hatred and the speech itself. For example, in the phrase pillow talk, the word pillow does not refer to the literal qualities of pillows such as softness or support but connotes the idea that pillows are typically found on beds, which is the same relation in which pillow talk occurs. It happens on or in beds. Similarly, the contribution of the word hate in hate speech could be to identify a relational structure that exists in both hate and hate speech.

This begs the question on the type and nature of this metaphoric relation. It largely depends on the meaning of hate, but a plausible interpretation is that hate or hatred is typically directed toward or against something or someone. Thus, hate in hate speech signifies being against something or someone, analogous to how hate operates. In its plain or literal sense, hate does not specify the object of hate. However, when combined with speech or crime, the term hate might not convey its plain or literal meaning but something more specific. As Cortese (2006) notes, hate has begun to be used in a more restricted sense since the mid-1980s to characterize negative beliefs and feelings about members of certain categories of people based on ethnicity, race, gender, sexual orientation, religion, age, or disability. This specialized usage continues in current discussions on hate speech and hate crimes.

Brown (2017a) suggests that the word hate can have different meanings depending on the context. Its core meaning involves intense dislike, aversion, or hostility toward something or someone, as seen in the term hate mail and the phrase haters gonna hate. Yet, in specialized contexts like hate speech and hate crime, hate can signify opposition to members of groups or classes identified by protected characteristics. Brown's (2017a) analysis reveals that hate speech cannot be fully understood through pure compositional semantics since hate assumes a figurative or metaphorical meaning. Therefore, the metaphoric meaning of hate speech is speech directed "against members of groups or classes of persons identified by protected characteristics analogous to how hate is toward or against something or someone" (Brown, 2017a, 464).

Brown (2017a) also contests the idea that the term hate speech possesses a singular, unambiguous meaning, which might eventually lead to a universally accepted definition reflecting this uniformity. They argue against the notion that hate speech is univocal, suggesting instead that it is equivocal, embodying a range of meanings rather than a single, precise one. According to Brown (2017b), hate speech represents a family of meanings, each contextually defined and understood, which precludes the possibility of pinning down an overarching, exact definition. Their perspective emphasizes the complexity and variability of how hate speech is perceived and used across different discussions, highlighting the challenges in establishing a universally agreed-upon definition.

In conclusion, the most important takeaway from examining hate speech through diverse perspectives is the recognition of its multifaceted nature. Rather than being driven solely by hatred, hate speech can stem from a range of motivations including fear, alienation, economic self-interest, and a desire for attention. This nuanced understanding challenges the simplification of hate speech as merely expressions of hatred, emphasizing the need for a broader, more context-sensitive approach. The final definition adopted in this thesis follows Brown's (2017a, 464): Hate speech is speech directed "against members of groups or classes of persons identified by protected characteristics analogous to how hate is toward or against something or someone".

## 2.2. The Impact of Hate Speech on Individuals and Society

As Lawrence (1990) suggests, the pain of being targeted by hate speech is not experienced by everyone, nor is the societal harm it causes borne equally. Often, there is a rush to claim understanding of the victims' suffering without truly listening, and a readiness to believe that one has endured the same. This can lead to the mistaken belief that balancing the protection of free speech with addressing the harm caused by hate speech is easy, resulting in the underestimation or misjudgment of the actual harm.

Recognizing these complexities, this subsection scrutinizes the harm inflicted by hate speech. It begins by examining the immediate harm inflicted on the direct targets of hate speech. Following this, the focus shifts to the broader impact on dominant non-target groups, highlighting how hate speech affects their perceptions and interactions. Finally, the discussion explores the societal implications of hate speech, illustrating its role in perpetuating social inequalities and inciting violence.

**Impact of Hate Speech on the Targets.** The examination of immediate harm begins with the most apparent victims of hate speech, namely those who are directly addressed. The links between verbal abuse and mental or emotional distress have been well established, challenging the old adage *sticks and stones may break my bones but words shall never hurt me*, which implies that verbal aggression is harmless compared to physical acts (Knoblock, 2022). Indeed, acute psychological or emotional turmoil emerges as the most direct harm caused by hate speech (Delgado, 1982). Words, whether filled with hate or not, can inflict mental, emotional, or even physical harm on their targets, especially if expressed in the presence of others or by someone in a position of authority. In fact, victims have described experiencing this type of harm in profoundly deep ways, akin to an existential kind of pain, as noted by Gelber and McNamara (2016). For example, victims may experience severely diminished self-esteem. According to Delgado (1982), the constant exposure to negative images forces upon the targets a harsh and destructive dilemma, either to despise oneself, as systematically demanded by society, or to forfeit any sense of self and become nothing. This suffering often intensifies due to a perception of the situation's hopelessness and even self-blame, leading to a cycle of self-reproach that exacerbates feelings of loneliness and undesirability. The psychological reactions to such stigmatization include feelings of humiliation, isolation, and self-loathing. Therefore, it is neither unusual nor pathological for stigmatized individuals to experience conflicted feelings regarding their self-worth and identity. This ambivalence stems from the stigmatized individual's awareness of being perceived as not meeting societal norms, which they have internalized. Consequently, stigmatized individuals often exhibit hypersensitivity and anticipate discomfort in interactions with those deemed normal and they frequently experience anxiety and fear, as detailed by Delgado (1982).

Beyond these immediate effects, the psychological impact of racism can manifest in mental illness and psychosomatic diseases, with affected individuals often seeking escape through substances like alcohol, drugs, or engaging in other forms of antisocial behavior as outlined by Delgado (1982). Importantly, achieving a high socioeconomic status does not mitigate the psychological damage inflicted by hate speech, as tragically demonstrated by the 2022 death of Dr. Lisa-Maria Kellermayr, an Austrian doctor who was harassed and threatened online for their advocacy of COVID-19 measures. The constant digital hate and threats ultimately led to their suicide (HateAid, 2022).

Delgado (1982) further stresses that the pursuit of success in business and managerial careers imposes a considerable psychological burden, even on those who are exceptionally ambitious and upwardly mobile within their identity groups. Moreover, individuals who achieve success often do not fully reap the benefits of their professional status due to inconsistent treatment by others, which leads to ongoing psychological stress, strain, and frustration. As a result, severe psychological impairments resulting from the environmental stress of prejudice and discrimination do not show a decreased incidence among minority group members with a higher socioeconomic status.

The stress associated with hate speech may manifest in physical health issues as well. Delgado (1982) highlights evidence suggesting that high blood pressure is linked with inhibited, constrained, or restricted anger rather than genetic factors, with hate speech contributing to elevated blood pressure levels.

Racial stigmatization can also adversely affect the target's financial interests. Indeed, the psychological injuries can significantly hinder an individual's career progression. Those who are timid, withdrawn, bitter, hypertensive, or psychotic are more likely to encounter difficulties in employment settings. Delgado (1982) references an experiment in which African American and white American of similar aptitudes were placed in a competitive environment, revealing that the African American participants exhibited defeatism, lackluster competitiveness, and a high expectancy of failure. For many minority group members, equalizing tangible variables such as salary and entry level is an insufficient remedy to counteract defeatist attitudes. The psychological cost of attempting to compete is too high, leading them to be "programmed for failure" (Delgado, 1982, 139-140).

**Impact of Hate Speech on Dominant Non-Target-Groups.** The impact of hate speech extends beyond its direct targets, potentially affecting non-target-group members in ways that often remain unnoticed. As noted by Matsuda (1989, 2338), members of dominant groups, who may vehemently oppose hate speech, find themselves harboring a "guilty secret". They may feel a sense of relief that they are not the subjects of racist attacks. Although they denounce groups like the Ku Klux Klan, there exists an ambivalent relief in not being in the targeted group, drawing them into an unwilling complacency with the perpetrators, grateful for not being the object of fear and degradation. This phenomenon is akin to the relief felt in the aftermath of human tragedies, such as natural disasters or plane crashes, where the fortunate distance themselves from the victims, thereby making it more challenging to foster a sense of common humanity. The prevalence of hate speech thus may create a rift between well-intentioned members of dominant groups and the victims, impeding empathy and solidarity.

Furthermore, Matsuda (1989) points out that hate speech prompts members of victimized groups to regard all dominant-group members with suspicion, while obliging well-meaning individuals from dominant groups to exercise excessive caution in their interactions with those considered outsiders. According to Matsuda (1989), despite efforts to resist, the notion of racial inferiority becomes subtly ingrained in people's minds as potentially valid due to the repetitive presentation of hate speech messages. Stereotypes that label the target group as lazy, dirty, sexualized, money-grubbing, or dishonest are vehemently rejected, yet they may surreptitiously influence perceptions and interactions. Matsuda (1989) argues that when in proximity to a member of the stigmatized group, the derogatory messages, regardless of conscious rejection, are involuntarily recalled, affecting behavior and connection with the individual.

For victims, the process involves a complex interplay of angry rejection and unintended absorption of the inferiority message, as described by Matsuda (1989). This dynamic highlights the pervasive and insidious effect of hate speech, not only reinforcing stereotypes among members of the dominant group but also compelling victims to contend with internalized messages of inferiority, further complicating the struggle for equality and mutual understanding.

Surprisingly, the perpetrator also emerges as a victim of hate speech. According to Delgado (1982), racial labeling and insults not only harm the targets but also inflict direct damage on the individuals who perpetrate such acts. Delgado (1982) suggests that bigotry could adversely affect those who harbor it by cementing rigid thought patterns, which in turn could dull their moral and social sensibilities and may lead to the development of a mentality that could be described as "mildly ... paranoid [sic]" (Delgado, 1982, 140). Contrary to some beliefs, there seems to be scant evidence to support the notion that racial slurs act as a "safety valve" for alleviating anxiety that might otherwise manifest in violent behavior, as Delgado (1982, 140) further notes.

**Impact of Hate Speech on Society.** Moreover, hate speech has the capacity to inflict harm not only directly, by eliciting fear, insecurity, and anxiety among its targets, but also in a somewhat indirect manner, by influencing the social hierarchy positions of the groups to which these targets belong, as discussed by Maitra and McGowan (2012). Specifically, speech can crystallize facts concerning the distribution of social power, outlining who possesses power and who does not.

For example, MacKinnon (1993) highlights how the imagery and language employed in pornography condition its viewers to become sexually aroused by the degradation of women, positing this conditioning as unconscious and therefore not subject to mental mediation (MacKinnon, 1993, 16). Also, Scoccia (1996) draws upon Austin's (1975) speech act theory to argue against pornography. They suggest that violent and certain nonviolent pornography executes a speech act with the illocutionary force of subordinating women and the perlocutionary effect of reinforcing women's subordinate sociopolitical status. To clarify, Austin's (1975) theory distinguishes two effects of speech. First, illocutionary force refers to the intended action resulting from the utterance. In Scoccia's (1996) argumentation, the production or dissemination of specific types of pornography aims, consciously or unconsciously, to position women in a subservient status relative to men. Second, perlocutionary force concerns the actual effects or outcomes precipitated by the utterance. In the context of the pornography that Scoccia (1996) addresses, it reinforces the societal placement of women as inferior.

Thus, the argument extends beyond the portrayal of women in submissive roles within pornography; it implicates such content in actively maintaining their subordinate status in reality, affecting both perceptions and treatments of women.

Langton (1993) supports this viewpoint, contending that speech has the power to subjugate by unjustly classifying women as inferior, stripping them of crucial rights and capabilities, and validating discriminatory treatment towards them. This line of reasoning underscores the profound and multifaceted impact of hate speech, particularly in contexts like pornography, on perpetuating and legitimizing social inequalities.

Speech can also inflict harm through prompting its listeners to emulate the behaviors presented. Empirical research highlights the human propensity for imitation, which is both strong and widespread (Hurley, 2004). This suggests that individuals may unconsciously replicate behaviors witnessed in mediums such as pornography, often without understanding the act of imitation or its motivations (Maitra and McGowan, 2012).

Furthermore, West (2012) study the detrimental effects of speech, particularly examining how certain hate speech can effectively silence individuals of color, thus infringing upon their right to free speech. They assert that to be deemed as truly embodying freedom of speech, such a right must meet three fundamental requirements: minimal distribution, minimal comprehension, and minimal consideration. Consequently, if racist hate speech obstructs the distribution, comprehension, or consideration of other speech, it can be seen as undermining rather than supporting the principle of free speech. Gelber and McNamara's (2016) empirical findings, drawn from interviews, lend support to West's (2012) argument, indicating that hate speech can disarm its targets, preventing them from acting against it. The interviews revealed that hate speech often leads to a withdrawal from public discourse, with some participants explicitly stating that hate speech left them unable to respond in the moment and silenced them in broader, more subtle ways. Similarly, some interviewees identified silence and withdrawal as strategies used by community members to shield themselves from hate speech.

**Hate Speech as a Precursor of Violence.** Research has consistently shown a correlation between verbal and physical violence, with verbal aggression often preceding and signaling the potential for physical abuse. Stets (1990) found that verbal aggression is a common precursor to physical acts of violence, indicating that verbal assaults should not be dismissed as benign. Similarly, Davis (1996) observed adults who made verbal threats towards children and noted that these threats often escalated to physical violence. Further exploring the dynamics within intimate relationships, Schumacher and Leonard (2005) identified prior verbal aggression by marital partners as significant predictors of physical aggression, highlighting the progression from verbal to physical abuse in domestic settings.

The potential harm of polarizing and hateful discourse extends beyond personal relationships into broader societal impacts. Leezenberg (2015) critically analyzes the rhetoric of Dutch politician Geert Wilders and the Freedom Party (PVV), illustrating how such discourse can polarize societies and contribute to violent acts, as seen in the tragic case of Anders Breivik's attack. On July 22, 2011, Anders Behring Breivik carried out a massacre in Norway, resulting in the deaths of 77 people and severely injuring 42 others. The attacker first set off a bomb in Oslo, then traveled to the island of Utøya, and opened fire on young people attending a youth camp (Leonard et al., 2014). Breivik's own admission of being partly inspired by Wilders's political rhetoric underscores the dangerous influence of hate speech on individuals predisposed to violence.

Additionally, the advent of computer-mediated communication has significantly increased the visibility and spread of discourse that was previously confined to private or marginal public spaces (Knoblock, 2022). According to Keipi et al. (2016), online platforms could facilitate the formation of extremist communities, providing a potential space for radicalized individuals to find validation and encouragement for their views. They suggest that the cases of Anders Breivik, various school shooters, and young ISIS recruits demonstrate the possible real-world consequences of hate speech's proliferation online.

United Nations experts (2019) highlighted how exposure to hate speech increases the likelihood of hate crimes, emphasizing the danger it poses to society. They noted that hate speech, common both online and offline, intensifies societal and racial tensions. This can lead to violent attacks and weaken democratic values, social stability, and peace. Moreover, Costello and Hawdon's (2018) observation that hate speech can create a snowball effect, where exposure leads to further production of hate material, highlights the urgent need for strategies to counteract this cycle and protect both individuals and the broader social fabric from its destructive impact.

In conclusion, hate speech inflicts severe psychological and emotional harm on its direct targets. Additionally, it disrupts social dynamics within dominant non-target groups, perpetuates societal inequalities, and incites violence.

## 2.3. The Debate Over Hate Speech and Free Speech Rights in Democratic Societies

In the previous subsection, the various impacts of hate speech on both direct targets and broader societal groups were explored. Given the severity of these consequences, the argument for protecting hate speech appears challenging, if not entirely misplaced. Indeed, Baker (2012), an advocate for the near-absolute protection of free speech, highlights a common perspective by recognizing the pervasive role that hate has played in the genocides and murderous racial conflicts of the twentieth century. The argument presented by Baker (2012) posits that the horrors of race-based violence diminish the significance of lesser values like free speech. Thus, from this viewpoint, it becomes difficult to justify the freedom to engage in hate speech when weighed against the potential to prevent racial violence and genocide. The proposition is that even if banning hate speech only potentially prevents such extreme acts, this possibility sufficiently justifies such legal restrictions.

However, there are compelling arguments against hate speech prohibitions that advocate for the preservation of free speech rights. Dworkin (2009) for example challenges the view that prohibiting hate speech is necessary to prevent discrimination and violence. They suggest that while it is crucial to legislate against discrimination and violence, restricting expressions of hate could compromise the legitimacy of these very laws. They argue that even if hate speech contributes to a hostile environment, curtailing it through legal means might not be justified if doing so undermines the democratic justification for other important laws. They emphasize that, "[w]e might have the power to silence those we despise, but it would be at the cost of political legitimacy, which is more important than they are"(Dworkin, 2009, ix).

Unlike Dworkin (2009), Waldron (2010) argues that in a well-ordered society, laws against hate speech are justified because they contribute to the dignity and mutual respect required by principles of justice. They believe such laws ensure public assurance of commitment to justice among citizens. Similarly to how societal order depends on laws against murder to convey public reassurance that murder is unacceptable, it also relies on prohibitions against hate speech to demonstrate that intolerance and bigotry are not condoned.

Baker (2012) counters that true assurance of mutual respect and dignity cannot be achieved through compelled speech conformity. Instead, it requires the freedom to express dissenting views, so society truly knows the beliefs and attitudes of its members. Baker (2012) argues against the facade of a well-ordered society enforced by coercive laws, comparing it to a Potemkin village which is neatly organized in appearance but devoid of genuine freedom. They further challenge the notion that the disciplinary role of the law is necessary or effective for fostering genuine respect and dignity among citizens. Drawing parallels with historical examples, such as compulsory flag salutes, Baker (2012) argues that such compulsion is often counterproductive and that fostering true loyalty and respect should be voluntary and spontaneous.

Baker (2012) further contends that the legitimacy of any legal order hinges upon its respect for individuals' equality and autonomy. They emphasize that true respect for autonomy means allowing individuals to express their values freely through speech, irrespective of the speech's content, of its potential negative impacts, or the difficulties it may create for government processes.

Similarly, Dworkin (2009) argues that allowing free expression, including hate speech, is a necessary condition for the legitimacy of laws in a democracy, particularly laws that are opposed by some segments of society. According to Dworkin (2009), every citizen must have not just a vote but also a voice. This implies that for a law to be considered legitimate, everyone must have had the opportunity to express their opinions freely, including those who oppose the law. This is crucial for maintaining the fairness of democratic decisions.

Dworkin (2009) emphasizes that the cultural and moral environment, which is shaped by the broad and unrestricted expression of ideas and attitudes, including prejudices and tastes, heavily influences legislation and policy. If certain views are suppressed, even if they are offensive or hateful, it undermines the fair democratic process and potentially delegitimizes the laws enacted in such an environment. After all, the law is considered legitimate precisely because it was established through the democratic process.

Dworkin (2009) further stresses the necessity and possibility of safeguarding women, homosexuals, and minority groups from the harmful effects of sexism, intolerance, and racism. They advocate for measures to prevent discrimination and inequality in various areas such as employment, education, housing, and the judicial system. Additionally, they support the enactment of laws to ensure their protection. But they further state that:

> [. . . ] we must not try to intervene further upstream, by forbidding any expression of the attitudes or prejudices that we think nourish such unfairness or inequality, because if we intervene too soon in the process through which collective opinion is formed, we spoil the only democratic justification we have for insisting that everyone obey these laws, even those who hate and resent them. (Dworkin, 2009, viii)

Likewise, Baker (2012) emphasize that the American conception of democracy relies heavily on individuals being able to form their political opinions autonomously. This indicates that individuals should be granted the liberty to explore, advocate, or listen to various views, including those that might contradict democratic principles, without legal restrictions on public discourse. The idea is that such unrestricted discourse is essential for people to authentically develop their own political commitments.

Even so, Waldron (2012) challenges Dworkin's (2009) and Baker's (2012) assertion that hate speech laws necessarily undermine the legitimacy of other democratic laws. They state that there are other ways for individuals to express their opposition to laws without resorting to hate speech and argues that if alternative forms of expression are available that do not involve hate speech, the legitimacy of the law is not compromised, because the essence of the opposition can still be expressed in a lawful manner.

Additionally, Baker (2012) makes an important distinction between formal and substantive autonomy. Formal autonomy is about the freedom to express one's values and make one's own choices, which should generally not be interfered with by the state. Substantive autonomy, on the other hand, relates to the actual capacity and opportunities individuals have to lead fulfilling lives, which might involve state intervention in providing resources or information. While laws can enhance substantive autonomy, they should not do so at the expense of violating formal autonomy. They argue that allowing one person to express their autonomy does not inherently conflict with another's autonomy. This is because formal autonomy does not involve controlling others or achieving specific outcomes, which would potentially lead to conflicts. In a democratic society, advancing people's substantive autonomy, like welfare or egalitarian aims, should not involve methods that disrespect individuals' formal autonomy. Particularly, Baker (2012) points out that typically racist hate speech, while repugnant and harmful in terms of undermining others' dignity, is an expression of the speaker's values and thus protected under the speaker's formal autonomy. Baker (2012) argues that legal prohibitions on hate speech generally should be impermissible as they violate the speaker's formal autonomy. Although, they note exceptions in specific institutional contexts where autonomy might be legitimately restricted, such as in employment settings where one might have to conform to role demands that preclude hate speech.

In contrast, Waldron (2012) acknowledges that while free speech is a foundational principle, there are recognized limits where speech can be regulated or criminalized, such as incitement to violence, fighting words, threats, obscenity, and defamation. These exceptions to free speech are based either on the immediate harm they cause or on their categorization as outside the protective umbrella of free speech due to their nature. Waldron (2012) presents two approaches to understanding exceptions to free speech:

- A balancing approach suggests that while free speech is important, it can be outweighed by other significant harms triggered by specific types of speech, such as hate speech.
- An intrinsic approach argues that some types of speech, like hate speech, are inherently outside the scope of free speech protection because they do not contribute to public discourse in a beneficial way and are aimed at harming others or undermining public order.

Waldron (2012) advocates for an honest recognition of the trade-offs involved in regulating hate speech. They criticize opponents of hate speech legislation for dismissing or downplaying the real harms caused by such speech and for overemphasizing the value of free speech without adequately considering its negative impacts.

While Baker (2012) acknowledges the harm caused by hate speech, they also question whether changing the approach to free speech could effectively prevent atrocities. Baker (2012) points out that while historical instances like the Holocaust or Rwandan genocide featured prominent racist hate speech, it is unclear whether such speech was a causal factor or merely symptomatic of deeper, underlying societal issues. They stress the importance of distinguishing between speech that directly causes harm and speech that simply reflects existing prejudices and social conditions and raises doubts about the effectiveness of legal prohibitions on hate speech.

Correspondingly, Knoblock (2022) criticizes the persistence of a simplified interpretation of linguistic relativity within both non-specialist and linguistic communities, noting how this notion periodically leads to calls for banning certain words with the expectation that this will resolve underlying social issues. The concept of linguistic relativity, also known as the Sapir-Whorf hypothesis (Benjamin Lee et al., 2012), posits that the structure of a language influences its speakers' worldview or cognition, thereby shaping their perceptions of the world. This theory frequently results in the belief that eliminating problematic words or grammatical features will consequently eradicate social ills, such as racism and sexism. However, practical examples and social reality urge caution against such optimistic beliefs. One illustrative case provided by Knoblock (2022) is the symbolic burial of the N-word in 2007, where thousands, including Detroit's mayor and Michigan's governor, gathered to ceremonially lay the word to rest in a coffin, complete with ribbons and roses, with the expectation that this act would contribute to ending racism. However, such symbolic gestures are insufficient as the systemic discrimination of African Americans remains a prevalent and persistent issue (Knoblock, 2022).

Baker (2012) continues by outlining three empirical conditions that would need to be met to justify hate speech regulations convincingly:

- Demonstrating that hate speech occurs in contexts leading to severe discrimination or genocide. Admittedly, this point can be demonstrated easily.

- Showing that the specific forms of hate speech that would be regulated are causally linked to these harms.

- Proving that legal prohibitions would effectively intervene in the causal chain to prevent these harms.

Finally, Baker (2012) explains why they believe that hate speech prohibitions are likely to backfire. First, Baker (2012) hypothesizes that hate speech prohibitions may not effectively reduce the likelihood of severe outcomes like genocide or systemic racism. They believe that simply outlawing hate speech does not address the underlying attitudes and societal conditions that lead to such outcomes. Second, Baker (2012) fears that these prohibitions could worsen the situation. By driving hate speech underground, such laws might prevent society from recognizing and addressing the true extent and nature of racist sentiments. Third, prohibitions might divert attention and resources away from more critical activities like openly confronting and refuting racist views. Baker (2012) warns that without the opportunity to counter misguided views openly, society risks turning truth into sterile dogma, losing the ability to effectively argue against harmful ideologies. Fourth, by removing overt expressions of racism from public discourse, society loses the opportunity to publicly reject these views, potentially leading to a failure to engage with and comprehend the reasons why these views are harmful. This could also create a platform for racists to claim victimhood, appealing to broader sentiments of free speech and liberty, thus potentially garnering unintended sympathy. Fifth, permitting hate speech within legal limits provides the benefit of identifying and understanding adversarial perspectives. Allowing such speech exposes the extent of racism, making it visible and identifiable. This visibility is crucial for those targeted by such speech to understand the threats they face and to organize effective counteractions. Lastly, prohibitions could intensify the sense of oppression among those with racist views, increasing their rage and conviction in their beliefs, and possibly leading to more extreme actions as a form of resistance against what they perceive as unjust suppression.

Significantly, Baker (2012) articulates concerns that a political agenda focused on enacting and enforcing hate speech prohibitions might inadvertently divert political energy from potentially more effective solutions to address the root causes of hate. They suggest that efforts could be more productively invested in three alternative strategies. First, improving the material conditions of identity groups who frequently become targets of hate speech could provide a more direct and impactful way to mitigate the effects of discrimination and bias. Second, a proactive and public expressive rejection of hate, often referred to as the "more speech" solution, can serve as a counterbalance to hate speech by promoting inclusive and affirming messages. Lastly, Baker (2012) suggests addressing the social circumstances that contribute to the development and spread of hateful attitudes and behaviors within identity groups. This involves combating issues such as social discrimination, and a sense of marginalization that can result in resentment and hostility.

In conclusion, the debate over hate speech and free speech rights in democratic societies reveals valid arguments on both sides. Advocates for banning hate speech emphasize the potential to prevent severe harms, such as racial violence and genocide, arguing that the societal benefits of such prohibitions outweigh the value of unrestricted free speech. Conversely, opponents of hate speech laws highlight the risks of undermining democratic legitimacy and political autonomy, asserting that free expression is crucial for genuine political discourse and societal transparency. Whether hate speech should be banned or not is beyond the scope of this thesis. However, the use of NLP methods to monitor hate speech presents a valuable tool for understanding and addressing its impact within democratic frameworks. Automated tools could provide a scalable and efficient means to monitor and manage the vast amounts of online communication, potentially identifying and mitigating hate speech before it escalates into real-world violence and discrimination. This technological approach does not replace the need for careful legal and ethical consideration but rather enhances our ability to uphold both safety and freedom in the digital age.

# 3. Machine Learning Preliminaries

This chapter introduces the machine learning approaches utilized during the experiments described in Chapter 5. First, logistic regression is discussed, followed by an examination of neural language models. Finally, the architecture of Transformer-based LLMs is addressed.

## 3.1. Logistic Regression

The following discussion on logistic regression is derived from Jurafsky and Martin's (2024b) comprehensive treatment of the topic in their work on speech and language processing.

A probabilistic machine learning classifier, such as logistic regression, consists of several key components (Jurafsky and Martin, 2024b):

1. Feature Representation of the Input: Each input observation $\mathbf{x^{(i)}}$ is represented as a vector of features. For example, if social media posts are classified as hate speech or other, features could be the frequency of specific words, the length of the post, or the presence of specific hashtags. Mathematically, this vector can be written as $[x_1, x_2, \dots, x_n]$, where each $x_i$ represents a specific feature of the observation.

2. Classification Function: The classification function calculates the estimated class $\hat{\mathbf{y}}$ based on the input features. This is typically done by computing the probability of each possible class given the input features. For binary logistic regression, the sigmoid function is used to convert the weighted sum of the features into a probability.

3. Objective Function for Learning: The objective function, also known as the loss function, evaluates how well the classifier performs and guides the learning process. It is designed to be minimized during training. In logistic regression, the cross-entropy loss function is frequently utilized. This function measures the difference between the predicted probabilities and the actual class labels.

4. Algorithm for Optimizing the Objective Function: An optimization algorithm is employed to refine the model parameters, i.e., weights and bias, in order to minimize the loss function. Gradient descent is a widely used optimization algorithm that updates the weights incrementally based on each training example.

Binary logistic regression is utilized to categorize data into one of two categories, such as hate speech or other. To classify a test instance, after learning the weights during training, the logistic regression model multiplies each feature $\mathbf{x_i}$ by its corresponding weight $\mathbf{w_i}$, sums these weighted features, and then adds a bias term $\mathbf{b}$. The value $\mathbf{z}$ denotes the sum of the weighted evidence for the class, as shown in Equation 1 (Jurafsky and Martin, 2024b, 3).

$$z = \left( \sum_{i=1}^{n} w_i x_i \right) + b \tag{1}$$

These parameters are utilized to calculate a score for each input observation. This score is subsequently transformed into a probability via the sigmoid function. If the probability is greater than 0.5, the observation is categorized as one class, e.g. hate speech, otherwise, it is categorized as the other class, e.g. other.

The sigmoid function is essential in logistic regression as it maps any real-valued number into the range (0, 1), making it suitable for probability estimation. As seen in Figure 1, outlier values get squashed toward 0 and 1 while the curve is nearly linear around 0. The function, defined as $\sigma(z) = \frac{1}{1+e^{-z}}$, converts the linear combination of input features and weights into a probability value (Jurafsky and Martin, 2024b, 3).

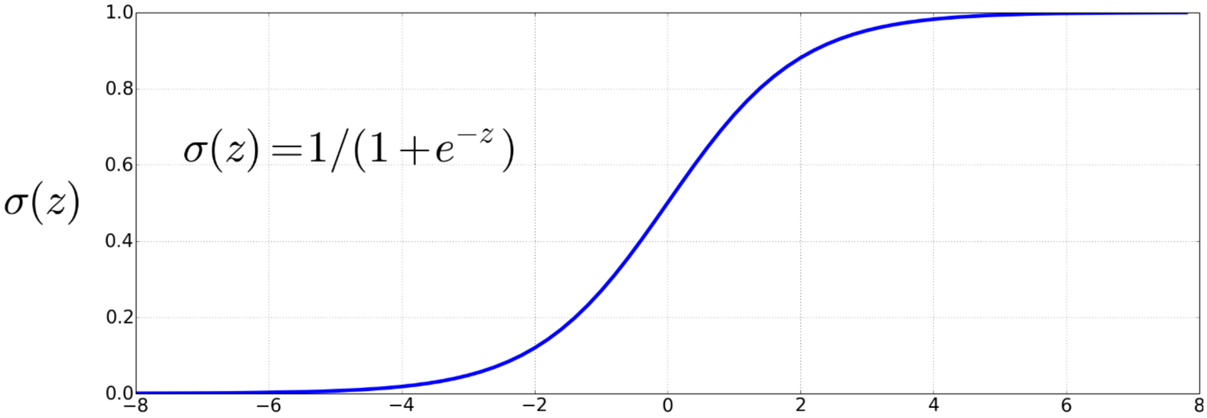


Figure 1.: Sigmoid Function Curve: Mapping Values to Probabilities (Jurafsky and Martin, 2024b, 3)

In logistic regression, the parameters of the model, specifically the weights $\mathbf{w}$ and the bias $\mathbf{b}$, are learned through a supervised classification process. This process involves using known correct labels $\mathbf{y}$, either 0 or 1, which could correspond to the classes hate speech or other, for each observation $\mathbf{x}$. The system generates an estimate of the true label $\mathbf{y}$, denoted as $\hat{\mathbf{y}}$, based on these parameters.

The goal is to find parameters $\mathbf{w}$ and $\mathbf{b}$ that approximate $\hat{\mathbf{y}}$ as close as possible to the true label $\mathbf{y}$ for each training observation. This requires two key components. The first is a metric to determine how close the current label estimate $\hat{\mathbf{y}}$ is to the true label $\mathbf{y}$. Rather than measuring similarity, this metric typically assesses the distance between the system's output and the true label, referred to as the previously mentioned loss function. The second component needed is an optimization algorithm that adjusts the weights iteratively in order to minimize the loss function. The standard algorithm for this purpose is stochastic gradient descent.

For each training example, the algorithm calculates the gradient of the loss function with respect to each parameter. The gradient indicates the direction and rate of change of the loss function. The parameters are subsequently adjusted in the opposite direction of the gradient to minimize the loss.

Figure 2 illustrates the gradient vector at the red dot in a two-dimensional space defined by $\mathbf{w}$ and $\mathbf{b}$. The red arrow in the $\mathbf{x}$-$\mathbf{y}$ plane points in the direction to move to find the minimum value, which is the opposite direction of the gradient. The gradient shows the direction of the increase, not the decrease.

The learning rate determines the size of the steps taken towards the minimum of the loss function. Selecting a suitable learning rate is important to ensure convergence. The

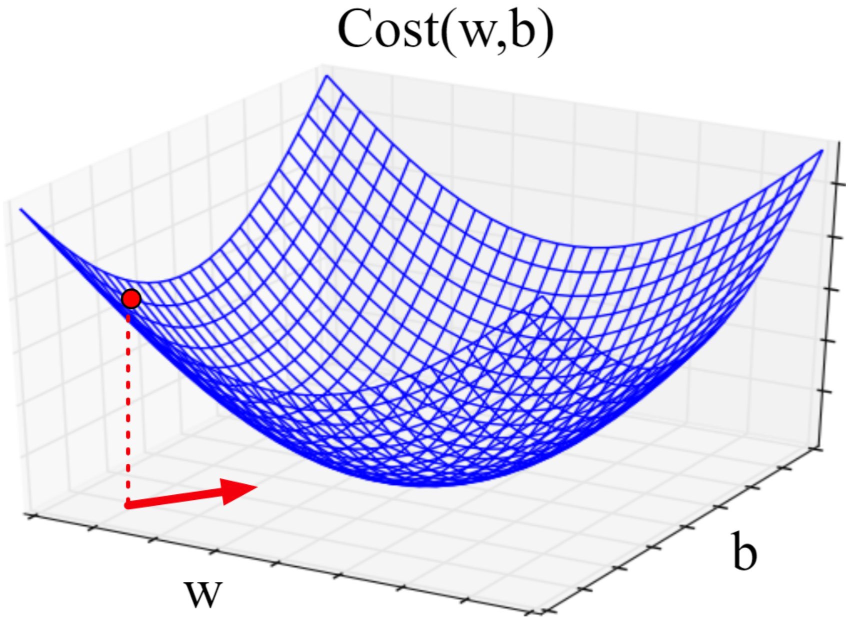


Figure 2.: Gradient Descent Optimization: Minimizing the Loss Function (Jurafsky and Martin, 2024b, 15)

learning rate is a small number that controls how big a step is taken when updating the parameters. If the learning rate is too high, the steps might be too big and the optimal values might be missed. If it is too low, the steps will be very small, and it will take a long time to get to the best values.

To understand logistic regression in the broader context of machine learning, it is helpful to compare it with other classification methods. Logistic regression belongs to discriminative models, which are designed to find the decision boundary that best separates different classes. Discriminative classifiers do not try to examine the characteristics of the classes themselves. Instead, they directly model the decision boundary between classes by learning which features best separate the classes. This approach focuses solely on learning to distinguish between different classes based on the input features without necessarily analyzing the underlying characteristics of the classes beyond their usefulness as discriminators. Conversely, generative classifiers, like Naive Bayes, focus on modeling the distribution of individual classes by analyzing the characteristics and features of each class. They aim to *generate* examples from the class distribution, and when given a new instance, they determine which class model best fits the instance.

To conclude, logistic regression stands as a robust supervised machine learning classifier that takes real-valued features, assigns each feature a weight, aggregates them, and processes the resultant sum through a sigmoid function to produce a probability. This probability is subsequently compared to a threshold to classify the input into a category. The model's parameters, comprising the weight vector **w** and bias **b**, are fine-tuned using a labeled training dataset. This process involves optimizing a loss function, commonly the cross-entropy loss, which measures the difference between predicted probabilities and actual class labels. The challenge of minimizing this loss function is tackled as a convex optimization problem, in which the objective is to find the parameter values that minimize the loss. To achieve this, iterative algorithms such as gradient descent are employed. These algorithms systematically adjust the weights to converge on the optimal values, thereby enhancing the model's predictive accuracy. Through this process of weight adjustment and probability estimation, logistic regression differentiates between classes based on the learned features.

## 3.2. Neural Language Models

The following section is derived from the work of Jurafsky and Martin (2024d), specifically their comprehensive coverage of neural language models and related neural network concepts.

Neural networks are an essential computational tool for NLP, originating from a simplified model of the biological neuron described in terms of propositional logic. Modern neural networks are networks of small computing units that take a vector of input values and produce a single output value, often applied to classification tasks.

Deep learning refers to the use of modern neural networks that are often deep, meaning they have many layers. This depth allows them to learn complex representations and features from raw data, making them powerful tools for tasks requiring significant data to learn features automatically.

Logistic regression and neural networks share much of the same mathematics, but neural networks are more powerful classifiers. A basic neural network with one hidden layer can learn any function, whereas logistic regression requires rich hand-derived features based on domain knowledge.

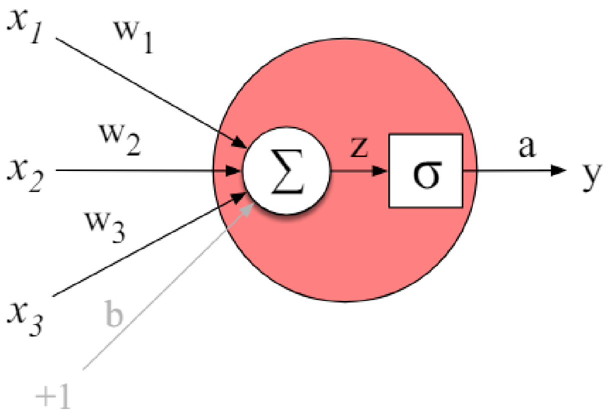


Figure 3.: The neural unit takes three input values $\mathbf{x_1}, \mathbf{x_2}$, and $\mathbf{x_3}$, multiplies each by a corresponding weight $\mathbf{w_1}, \mathbf{w_2}$, and $\mathbf{w_3}$, adds a bias term $\mathbf{b}$, and applies a sigmoid function, resulting in a value between 0 and 1. $\mathbf{y}$ denotes the final output, and $\mathbf{a}$ denotes the activation of an individual node (Jurafsky and Martin, 2024d, 3).

The building blocks of neural networks include units that perform computations on input values to produce an output. A bias term is an additional term in the weighted sum that a neural unit computes. A vector is a list or array of numbers used to represent inputs or weights. Activation refers to the non-linear function applied to the weighted sum, producing the final neuron's output. Frequently used activation functions include the sigmoid function, which maps output to the range (0,1), the tanh, or hyperbolic tangent function, which ranges from -1 to +1, and the ReLU, or rectified linear unit, which outputs the input if it is positive and zero otherwise. Figure 3 (Jurafsky and Martin, 2024d, 3) illustrates a basic neural unit. In this instance, the neural unit takes three input values $\mathbf{x_1}, \mathbf{x_2}$, and $\mathbf{x_3}$, calculates a weighted sum by multiplying each input by a corresponding weight, $\mathbf{w_1}, \mathbf{w_2}$, and $\mathbf{w_3}$, and adds a bias term $\mathbf{b}$. It subsequently applies a sigmoid function to the sum, resulting in a value between 0 and 1. $\mathbf{y}$ refers to the final output of the entire network and $\mathbf{a}$ means the output of the activation of an individual node.

A feedforward neural network is a type of artificial neural network where the connections between the nodes do not form a cycle. This means that the information moves in one direction. It flows from the input nodes, through the hidden nodes, and ultimately to the output nodes. Each layer in a feedforward network is fully connected to the next layer, meaning each node in one layer is connected to every node in the subsequent layer. This architecture is called feedforward because the data passes through the network in a single direction, without any feedback loops or convolutions.

A fully-connected feedforward neural network is a specific type of feedforward neural network in which each neuron in one layer is connected to every neuron in the next layer. This network typically consists of an input layer, one or more hidden layers, and an output layer. The computation in a fully-connected feedforward neural network involves three main steps. The first step is multiplying the weight matrix with the input vector $\mathbf{x}$. The second step is adding the bias vector $\mathbf{b}$. The final step is applying an activation function $\mathbf{g}$, such as the sigmoid, tanh, or ReLU, to produce the output of the network. The output of each layer becomes the input to the next layer, and this process continues until the final output layer is reached. A normalizing step like the softmax function is then needed to convert the output vector into a probability distribution, where all values lie between 0 and 1 and sum to 1, making the outputs interpretable as probabilities.

A 2-layer net can be represented as shown in Equation 2 (Jurafsky and Martin, 2024d, 10).

$$\begin{aligned} z^{[1]} &= \mathbf{W}^{[1]}a^{[0]} + \mathbf{b}^{[1]} \\ a^{[1]} &= g^{[1]}(z^{[1]}) \\ z^{[2]} &= \mathbf{W}^{[2]}a^{[1]} + \mathbf{b}^{[2]} \\ a^{[2]} &= g^{[2]}(z^{[2]}) \\ \hat{y} &= a^{[2]} \end{aligned} \tag{2}$$

The square-bracketed superscripts indicate the layer numbers, starting at 0 for the input layer. Therefore, $\mathbf{W}^{[1]}$ denotes the weight matrix for the first hidden layer, and $\mathbf{b}^{[1]}$ represents the bias vector for the first hidden layer. The function $\mathbf{g}(\cdot)$ refers to the activation function, which is typically ReLU or tanh for intermediate layers and softmax for output layers. The term $\mathbf{a^{[i]}}$ denotes the output from layer $\mathbf{i}$, and $\mathbf{z^{[i]}}$ is the combination of weights and biases $\mathbf{W}^{[i]}a^{[i-1]} + \mathbf{b}^{[i]}$. The 0th layer is for inputs, so the inputs $\mathbf{x}$ are generally referred to as $\mathbf{a^{[0]}}$ and $\hat{\mathbf{y}}$ denotes the current label estimate.

Thus, the algorithm for computing the forward step in an n-layer feedforward network involves iterating through each layer, computing the weighted sum and incorporating the bias, followed by the application of the activation function to produce the output of that layer. This logic is illustrated in Algorithm 1 (Jurafsky and Martin, 2024d, 10).

For NLP tasks, the input representation of a feedforward network often involves word embeddings, which are dense vector representations of words that capture their meanings and relationships. Pre-training refers to the use of embeddings learned from large datasets before using them for specific tasks. This process helps the network start with a good representation of words, improving its performance on the task at hand.

Neural networks are trained through the process of gradient descent optimization,

**Algorithm 1** Feedforward Neural Network Forward Pass

```
for $i$ in $1, \ldots, n$ do
    $z^{[i]} \leftarrow W^{[i]} a^{[i-1]} + b^{[i]}$
    $a^{[i]} \leftarrow g^{[i]}(z^{[i]})$
end for
$\hat{y} \leftarrow a^{[n]}$
```

where the network's parameters, i.e., weights and biases, are adjusted to minimize the difference between the predicted output and the true output. This training process involves computing the gradient of the loss function with respect to the network's parameters and updating the parameters accordingly. Error backpropagation is an algorithm employed to calculate the gradient of the loss function in relation to each parameter in the network. It entails a forward pass to determine the output and a backward pass to propagate the error and compute the gradients.

Neural language models are neural networks designed to predict the probability of a word sequence and upcoming words. They work by taking a context of previous words, represented by their embeddings, and predicting the next word in the sequence.

The classes to be predicted in neural language models are the words in the vocabulary. A language model acts as a word predictor by assigning probabilities to each word in the vocabulary based on the context provided by previous words.

Each of the $\mathbf{N}$ previous words is represented as a one-hot vector of length $|\mathbf{V}|$, with one dimension for each word in the vocabulary. A one-hot vector has one element equal to 1 in the dimension corresponding to that word's index in the vocabulary, while all other elements are set to zero. For example, in a one-hot representation for the word toothpaste, if the index is 5 in the vocabulary, $\mathbf{x_5 = 1}$ and $\mathbf{x_i = 0}$ for all $\mathbf{i \neq 5}$.

The feedforward neural language model, as illustrated in Figure 4, uses a a sliding window capable of viewing $\mathbf{n}$ words into the past. If $\mathbf{n}$ is set to 3, the three words $\mathbf{w_{t-1}}, \mathbf{w_{t-2}}$, and $\mathbf{w_{t-3}}$ are each represented as a one-hot vector. These one-hot vectors are then multiplied by the embedding matrix $\mathbf{E}$. The embedding weight matrix $\mathbf{E}$ includes a column for every word, with each column vector having $\mathbf{d}$ dimensions. As a result, $\mathbf{E}$ has dimensions of $\mathbf{d} \times |\mathbf{V}|$. By multiplying the embedding matrix $\mathbf{E}$ by a one-hot vector, the relevant column vector corresponding to the word is selected, yielding the word's embedding.

The three resulting embedding vectors are concatenated to produce the embedding layer $\mathbf{e}$. This embedding layer is succeeded by a hidden layer and an output layer, where the softmax function produces a probability distribution over words. For instance, $\mathbf{y_{42}}$, the value of output node 42, indicates the probability that the next word $\mathbf{w_t}$ is $\mathbf{V_{42}}$, the vocabulary word at index 42, which in this case is the word fish.

As shown in Figure 4, at each timestep $\mathbf{t}$, the network generates a $\mathbf{d}$-dimensional embedding for every context word by multiplying a one-hot vector with the embedding matrix $\mathbf{E}$. The resulting three embeddings are combined to create the embedding layer $\mathbf{e}$. This layer is then multiplied by a weight matrix $\mathbf{W}$, and an activation function is applied element-wise to form the hidden layer $\mathbf{h}$. The hidden layer $\mathbf{h}$ is subsequently multiplied by a different weight matrix $\mathbf{U}$. Finally, a softmax output layer estimates the probability at each node $\mathbf{i}$ that the next word $\mathbf{w_t}$ will correspond to the vocabulary word $\mathbf{V_i}$. In essence, the neural language model's equations with a window size of 3, given one-hot input vectors for each context word can be represented as in Equation 3 (Jurafsky and

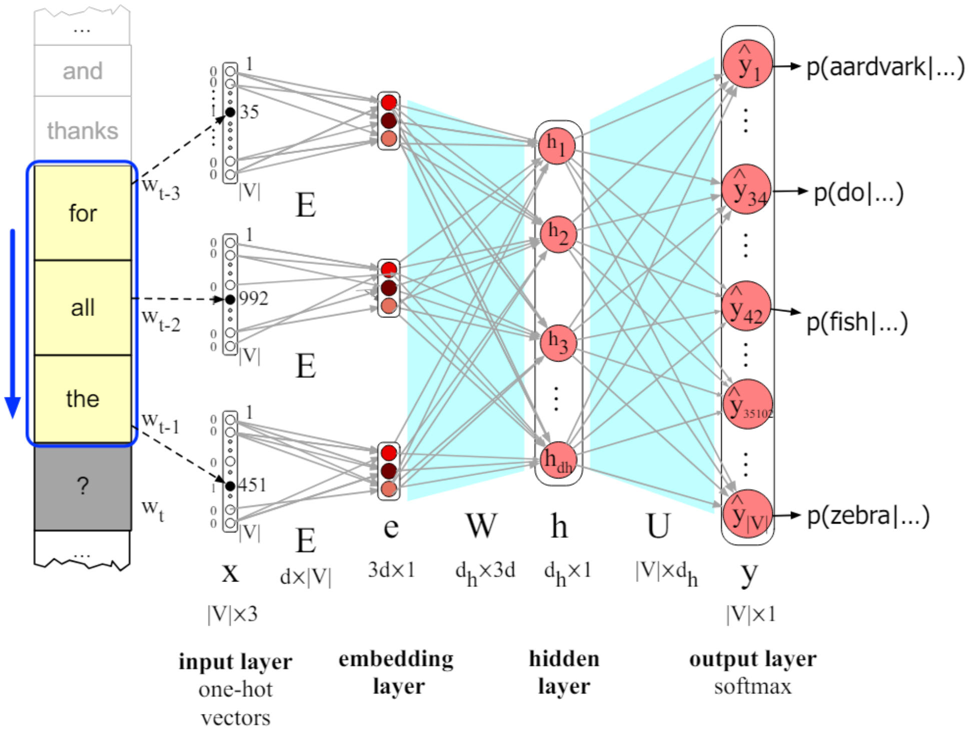


Figure 4.: Decoding or Forward Inference for Neural Language Models (Jurafsky and Martin, 2024d, 24)

Martin, 2024d, 23).

$$
\begin{aligned}
\mathbf{e} &= [\mathbf{E}x_{t-3}; \mathbf{E}x_{t-2}; \mathbf{E}x_{t-1}] \\
\mathbf{h} &= \sigma(\mathbf{W}\mathbf{e} + \mathbf{b}) \\
\mathbf{z} &= \mathbf{U}\mathbf{h} \\
\hat{\mathbf{y}} &= \text{softmax}(\mathbf{z})
\end{aligned} \tag{3}
$$

Neural language models are trained using self-supervised training, where the model is trained on a large corpus of text to predict the next word at each time step. This process does not require labeled data, as the sequence of words itself provides the supervision needed for training.

To conclude, neural networks consist of neural units, which are abstract computational elements inspired by biological neurons. Each unit processes input values by applying a weight vector, adding a bias, and using a non-linear activation function, such as ReLU, tanh, or sigmoid. In a fully-connected, feedforward network, every unit in one layer is connected to every unit in the next layer, forming an acyclic structure. The power of neural networks comes from their ability to let early layers learn representations that improve the performance of later layers. These networks are trained using optimization methods such as gradient descent. The backpropagation algorithm, which operates on a computation graph, is essential for calculating the gradients of the loss function to adjust the network parameters. Neural language models, a specific type of neural network, function as probabilistic classifiers in two main categories: discriminative or masked language models, such as BERT, which estimate the probability of missing words in a sequence based on the surrounding context, and generative language models, such as GPT, which predict entire sequences of words based on the preceding context. These models can use pre-trained embeddings or develop new embeddings during the language modeling process.

## 3.3. Transformer-Based LLMs

This section explores the components and functioning of Transformers (Vaswani et al., 2017), the foundational architecture behind LLMs, such as the GPT models used in this research. It begins by examining the architecture of GPT models, discussing the mechanics of self-attention, multihead attention, and the hierarchical structure of Transformer layers. Next, the focus shifts to different decoding techniques used in Transformer-based models for NLP tasks. Following this, the section examines how Transformers process input sequences, the role of embeddings, and the impact of model size and training data diversity on the performance of LLMs. Subsequently, the section compares discriminative and generative pre-trained Transformers. Finally, it discusses transfer learning methods. This section is informed by Jurafsky and Martin's (2024f) extensive analysis of the subject of Transformer-based LLMs in their work on speech and language processing.

**Components of Transformers.** LLMs are pre-trained language models that exhibit strong performance on various natural language processing tasks due to the extensive knowledge they acquire during pre-training. These models learn from vast amounts of text, enabling them to perform tasks such as summarization, machine translation, question answering, and chatbot interactions effectively.

The intuition behind Transformers, the fundamental architecture used for LLMs, is to build increasingly nuanced contextualized representations of input word meanings across a series of layers. At each layer, the Transformer combines information from the previous layer's representation of a word with information from the representations of neighboring words. This mechanism enables the model to generate a contextualized representation for each word at each position, helping to integrate the meaning of words based on their context.

Self-attention, a key mechanism in Transformers, allows the model to look broadly in the context and tells it how to integrate representations from words in that context to build a new representation for the current word. By computing the attention scores between words, the model can target pertinent parts of the context, even if those parts are far away from the current word. This process enables the model to identify complex relationships and dependencies within the text.

The fundamental concept of attention involves comparing an item of interest to a set of other items to determine their relevance within the current context. In self-attention for language, the set of comparisons is to other words or tokens within a given sequence. The outcomes of these comparisons are then utilized to generate an output sequence for the given input sequence. In self-attention, this process involves creating a set of query, key, and value vectors from the input embeddings. These vectors are used to compute scores between different positions in the input, which are subsequently normalized with the softmax function to yield attention weights. These weights are used to compute a weighted sum of the value vectors, producing the output for each position in the sequence.

The query, key, and value are vectors derived from the input embeddings by multiplying the input by learned weight matrices. The query vector indicates the current focus of attention when compared to all prior inputs. The key vector represents the preceding input being compared to the current focus of attention. The value vector is utilized to calculate the output for the current point of focus. These vectors are essential for computing the scores that determine how much attention each input receives during the self-attention process.

The self-attention process can be parallelized because each output is computed independently of all other outputs. This independence allows the entire process to utilize efficient matrix multiplication routines by consolidating the input embeddings of the tokens into a single matrix. This parallelization is significant because it enables the Transformer to process long sequences of input tokens more efficiently than recurrent neural networks, which process tokens sequentially.

Multihead self-attention layers are composed of several self-attention layers, known as heads, which operate in parallel at the same depth within a model, each possessing its own set of parameters. By using distinct sets of parameters, each head can learn different aspects of the relationships among inputs at the same level of abstraction. Multihead self-attention layers address the issue that a single self-attention model may struggle to capture all the different kinds of parallel relations among input words. Words in a sentence can simultaneously exhibit multiple types of relationships, such as syntactic, semantic, and discourse connections.

Transformers consist of several layers of Transformer blocks. A Transformer block is a multilayer network that converts sequences of input vectors into sequences of output vectors of the same length. It combines feedforward networks, residual connections, normalizing layers, and self-attention layers to extract and use information from large contexts.

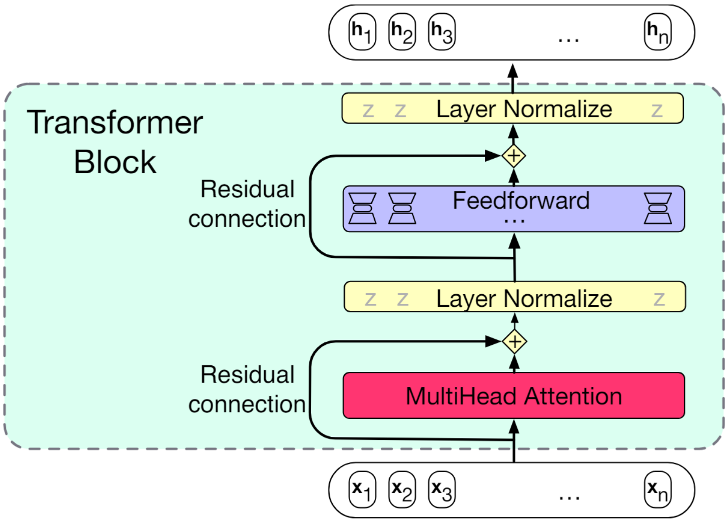


Figure 5.: Layers of a Transformer Block (Jurafsky and Martin, 2024f, 10)

The feedforward layer in a Transformer block contains position-wise networks, which are fully connected two-layer networks with one hidden layer. These networks are independent for each position, allowing parallel computation. Residual connections transfer information from a lower layer to an upper layer without passing through the intermediate layer, improving learning by giving higher-level layers direct access to information from lower layers. Layer normalization maintains the values of a hidden layer within a range that aids gradient-based training. It normalizes the summed vectors using the mean and standard deviation of the elements, followed by applying learnable parameters for gain and offset.

These components are put together in a sequence where the input undergoes self-attention, residual addition, layer normalization, feedforward processing, another residual addition, and a final layer normalization, as seen in Figure 5 (Jurafsky and Martin, 2024f, 10).

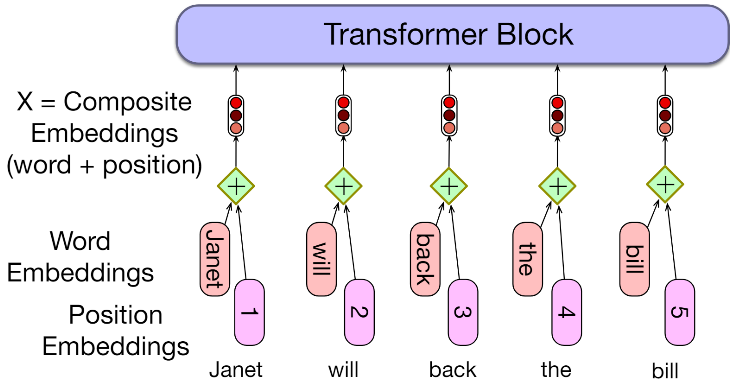


Figure 6.: Combining an embedding of the absolute position with the token embedding creates a new embedding of the same dimensionality (Jurafsky and Martin, 2024f, 16).

Next the origin of the input for the Transformers is discussed. As illustrated in Figure 6, the Transformer obtains its input by separately computing two types of embeddings: an input token embedding and an input positional embedding. Token embeddings are vectors representing the input tokens, while positional embeddings indicate the position of each token within the sequence. These embeddings are combined to form the input matrix for the Transformer model, capturing both the identity and the position of each token.

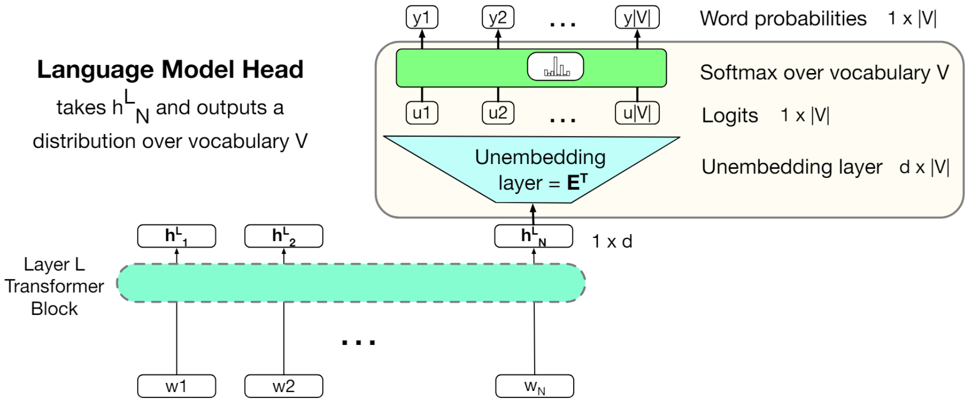

Language Model Head
takes $h^L_N$ and outputs a distribution over vocabulary V
y1
y2
...
y|V|
Word probabilities 1 x |V|
Softmax over vocabulary V
u1
u2
...
u|V|
Logits 1 x |V|
Unembedding layer = $E^T$
Unembedding layer d x |V|
$h^L_1$
$h^L_2$
$h^L_N$
1 x d
Layer L Transformer Block
...
w1
w2
$w_N$

Figure 7.: The language model head takes the output from the final Transformer layer and predicts the next word by computing a probability distribution across the complete vocabulary (Jurafsky and Martin, 2024f, 17).

The final element of the Transformers to be discussed is the language model head. The language model head is the component added on top of the Transformer to enable language modeling. Figure 7 shows how the language model head takes the output from the final Transformer layer and predicts the next word by computing a probability distribution across the complete vocabulary. This is achieved by a linear layer projecting the final layer's output to a logit vector, followed by a softmax layer to produce the probability distribution.

Figure 8 shows the complete architecture of a Transformer decoder-only model. The Transformer constructs layers of Transformer blocks to process a sequence of input tokens $w_1$ to $w_N$, ultimately predicting the next word $w_{N+1}$.

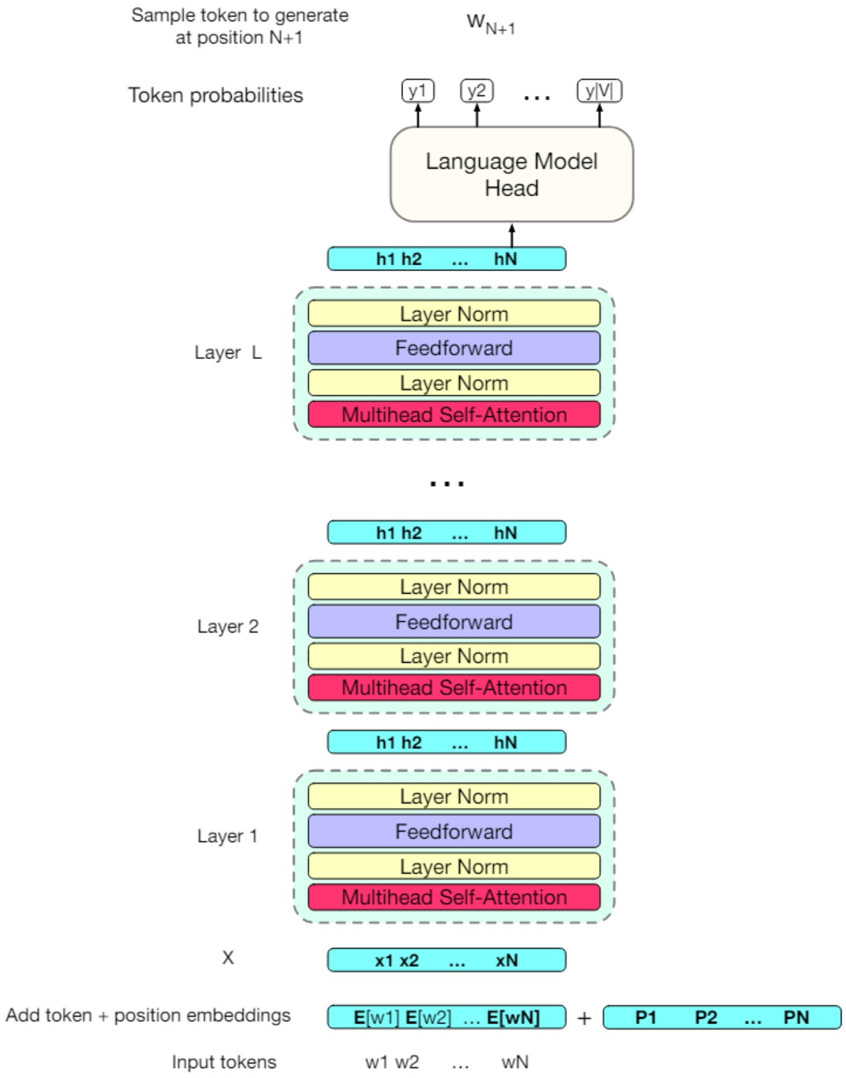


Figure 8.: Architecture of a Transformer Decoder-Only Model (Jurafsky and Martin, 2024f, 18)

**Decoding Techniques for Transformer-Based LLMs in NLP Tasks.** Transformer-based LLMs are applied to NLP tasks because they can capture contextual information and dependencies across large spans of text. Their ability to generate contextualized representations makes them highly effective for various tasks such as text generation, summarization, and translation. Their capabilities at predicting upcoming words is crucial because many practical NLP tasks can be cast as word prediction. A powerful language model can solve these tasks with high accuracy by generating appropriate text continuations based on the context provided. To apply Transformer-based LLMs to NLP tasks, the model is given a text prefix and asked to generate a possible completion. The model has access to the priming context and its own generated outputs within the large context window, allowing it to incorporate extensive contextual information at each step of the generation process.

This raises the question of how words are generated at each step. In the broader context of language models, decoding refers to the process of generating a sequence of words from a model's probability distributions. It involves selecting the next word in a sequence based on the probabilities assigned by the model, and it can be performed using various methods such as beam search, sampling, or greedy decoding.

Greedy decoding is one method of generating text where at each time step, the output word is selected by calculating the probability for each possible output and then choosing the word with the highest probability. This method can be problematic as it frequently results in generating text that is predictable, generic, and repetitive. Additionally, it is deterministic, meaning that the same input context will always produce the same output. To address these issues, more sophisticated decoding methods are often employed.

One such method is beam search, a decoding algorithm that maintains multiple hypotheses at each time step and explores multiple paths simultaneously to find a sequence of words that has the highest overall probability. By considering a wider range of possible sequences, beam search helps to mitigate the limitations of greedy decoding, potentially producing more accurate and coherent results. Beam search is particularly effective for tasks like machine translation, where text generation is highly constrained by a corresponding text in another language. However, for other NLP tasks, more advanced techniques known as autoregressive generation or sampling methods are favored, as they introduce more diversity into the generated text.

Autoregressive generation is a specific type of decoding where the model produces text word by word, with each word conditioned on the previously generated words. This left-to-right approach ensures that each word is generated based on the context of all preceding words, allowing the model to maintain coherence and context throughout the text generation process.

Sampling is another method of generating text, where words are chosen randomly according to their probability distribution, as predicted by the model. This approach adds an element of randomness to the generation process, potentially producing more diverse and creative outputs. When used in Transformers, sampling involves selecting words based on the model's predicted probabilities rather than always choosing the most probable word, which can lead to more varied text. The introduced sampling methods aim to strike a balance between quality and diversity. Methods that prioritize high-probability words tend to produce high-quality but less diverse text, while those that allow for more middle-probability words can produce more diverse but potentially lower-quality text.

To achieve this balance, various sampling techniques are employed. Top-k sampling is one such technique, where the model truncates the probability distribution to the top **k** most likely words and then randomly samples from these words. But this approach can be problematic because the fixed value of **k** might not adapt well to different contexts, leading to less flexible and sometimes less coherent outputs.

To overcome the limitations of top-k sampling, top-p sampling, or nucleus sampling, is used. This method keeps the top **p** percent of the probability mass rather than a fixed number of words, dynamically adjusting the number of candidate words based on their probabilities. This approach provides a more contextually appropriate balance between quality and diversity, making it more effective in various scenarios.

Another approach to refining the sampling process is temperature sampling, which reshapes the probability distribution by dividing the logits by a temperature parameter before applying the softmax function. The intuition behind temperature sampling is drawn from thermodynamics, where higher temperatures allow for greater flexibility and exploration of states, while lower temperatures restrict this exploration to a subset of better states. In low-temperature sampling, a smaller temperature value increases the probability of the most likely words and decreases the probability of less likely words, making the distribution more focused and greedier. Conversely, high-temperature sampling, with a larger temperature value, flattens the probability distribution, allowing for greater diversity in the generated text.

**Training Transformers for NLP.** To teach a Transformer to be a language model, the training process involves self-supervision using a corpus of text. At each time step, the model is instructed to predict the next word in the sequence. This does not require any special labels as the sequence of words provides its own supervision. The model is trained to reduce the error in predicting the true next word using cross-entropy as the loss function. This approach utilizes the inherent structure of language data to train the model effectively.

LLMs achieve their capabilities by being trained on vast amounts of text data. This extensive training enables them to learn the properties and patterns of language, allowing them to perform various NLP tasks effectively. The diversity and volume of the training data play a crucial role in the model's ability to generalize across different contexts and tasks.

The performance of LLMs is influenced by several key factors, including the size of the model, the amount and diversity of the training data, and the computational resources available for training. Larger models and more diverse datasets typically result in better performance, as they can identify more complex patterns and nuances in the data.

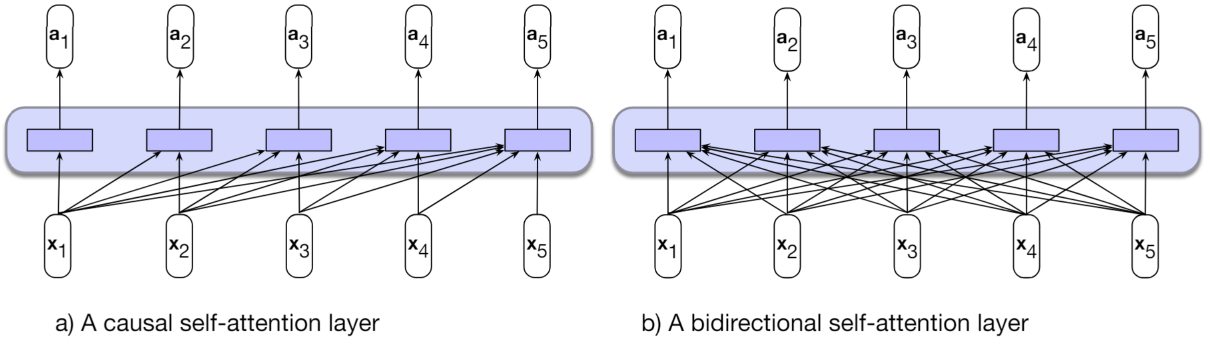
In summary, Transformers are advanced non-recurrent networks that utilize self-attention mechanisms. These mechanisms map input sequences to output sequences of the same length by using attention heads that assess the relevance of surrounding words to the current word. Each Transformer block contains one attention layer followed by a feedforward layer, enhanced with residual connections and layer normalizations. Stacking multiple Transformer blocks creates deeper and more capable networks. Constructing language models involves stacking Transformer blocks and adding a linear layer and a softmax layer at the top. Transformers feature a wide context window allowing them to use extensive context for predicting future words. This broad context makes Transformer-based models particularly effective for numerous NLP tasks by framing them as word prediction tasks. Additionally, the generation of words in LLMs is typically guided by a sampling algorithm.

**Discriminative vs. Generative Pre-Trained Language Models.** To better understand the functioning of GPT models, it is helpful to contrast them with their counterparts within the broader context of machine learning, namely the discriminative pre-trained models, such as Bidirectional BERT (Devlin et al., 2019). GPT models belong to the category of causal or generative pre-trained models, while BERT is a discriminative and masked pre-trained model.

As discussed in Section 3.1, generative and discriminative classifiers represent two fundamentally different approaches to machine learning. Generative classifiers, like Naive Bayes, focus on modeling the distribution of individual classes by analyzing the characteristics and features of each class. They aim to generate examples from the class distribution and, when given a new instance, determine which class model best fits the instance. In contrast, discriminative classifiers, such as logistic regression, do not try to examine the characteristics of the classes themselves. Instead, they directly model the decision boundary between classes by learning which features best separate the classes.

This distinction between generative and discriminative approaches extends beyond traditional classifiers to pre-trained language models as well. Discriminative pre-trained language models and generative or causal pre-trained language models differ in several ways. Discriminative models are created to assign input data to specific predefined categories. They learn the boundaries between different classes based on the input data. Conversely, generative models are designed to learn the probability distribution of sequences of words and can sample from this distribution to produce new sequences. Discriminative models select a category based on the likelihood given the entire input, while generative models are designed to generate new data points from the learned data distribution and can create new instances of data similar to the training data.

In terms of architecture, discriminative models such as BERT use a bidirectional attention mechanism, allowing them to consider both the left and the right context when making predictions, as illustrated in Figure 9. BERT is an encoder-only type of Transformer, meaning it utilizes only the encoder part of the Transformer architecture and processes the entire input sequence simultaneously. It uses the bidirectional self-attention mechanism to understand the context of each word in relation to both its preceding and succeeding words. During pre-training, BERT masks a percentage of the input tokens. The model is thus trained to predict the tokens that are masked based on the entire context, including both preceding and following words. Conversely, causal or generative models like GPT use a unidirectional, left-to-right attention mechanism. GPT is a decoder-only model, which means it utilizes only the decoder part of the Transformer architecture and generates each word in a sequence one at a time, only considering the preceding words and not the future words. This setup is essential for generating text because it enables the model to forecast the subsequent word in a sequence based on the previous context. In GPT models, the self-attention mechanism includes causal masking to stop the model from seeing upcoming tokens, ensuring that each word is predicted based only on the words that come before it.



$a_1$
$a_2$
$a_3$
$a_4$
$a_5$
$x_1$
$x_2$
$x_3$
$x_4$
$x_5$
a) A causal self-attention layer
b) A bidirectional self-attention layer

Figure 9.: (a) The causal or generative Transformer model, which looks backward, computes each output independently using only prior context information. (b) In the bidirectional self-attention model, each element of the sequence is processed by considering all inputs, both preceding and succeeding the current one (Jurafsky and Martin, 2024a, 2).

BERT's pre-training objective, masked language modeling, involves predicting randomly masked words within a sentence. GPT models, on the other hand, are trained with the objective of predicting the next word in a sequence. This characteristic makes them generative because they can produce new sequences of text by sequentially predicting each next word based on the context provided by previous words.

After pre-training, BERT is typically fine-tuned for specific downstream tasks such as text classification, question answering, and named entity recognition. These tasks are inherently discriminative because they involve classifying or labeling the input text rather than generating new text. The classifier head in BERT is designed for specific downstream tasks such as text classification, named entity recognition, and question answering. This head is typically a simple feedforward neural network or a logistic regression layer put on top of the BERT embeddings. During fine-tuning, the classifier head takes the contextualized embeddings produced by the BERT encoder, especially the [CLS] token representation, and outputs probabilities for the predefined categories. The classifier head makes BERT a discriminative model because it focuses on distinguishing between different categories or labels given the entire input. The model is trained to maximize the likelihood of the correct label, which is a typical characteristic of discriminative tasks.

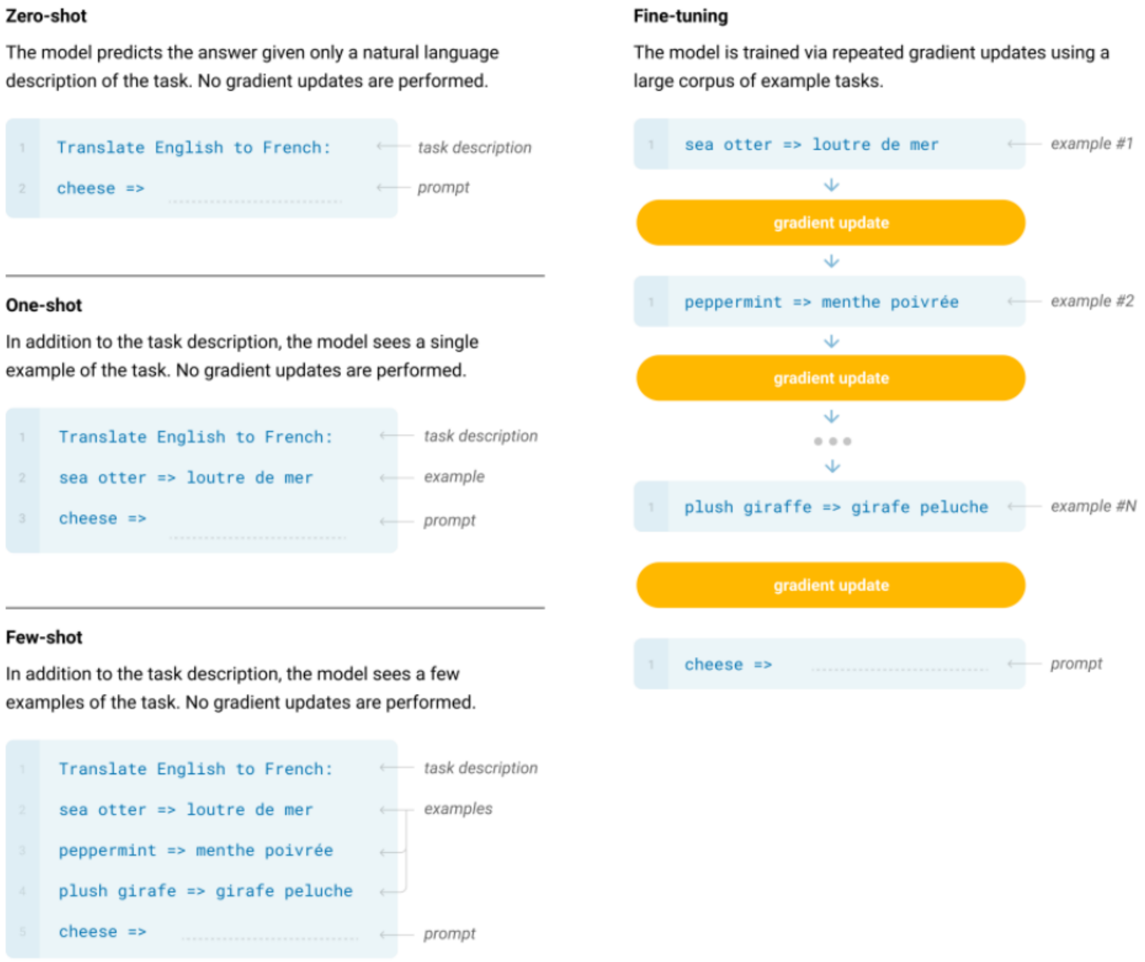
Conversely, the language model head in GPT is designed for generating text. This head outputs likelihoods for the subsequent word in the sequence, given the previous words. During text generation, the language model head uses the contextual embeddings from the GPT decoder to predict the next token. This is done iteratively, generating one word at a time based on the preceding context. The language model head makes GPT a generative model because it focuses on creating new sequences of text. The model is trained to maximize the likelihood of the next word in the sequence, which is the primary task of generative models.

**Transfer Learning and Prompting.** Transfer Learning involves using a model pre-trained on one task or domain to enhance performance on a different but related task or domain (Raffel et al., 2020). This leverages the knowledge the model has acquired from its initial training to perform well on new tasks with limited additional training. The experiments described in Section 5 utilized two such methods, namely zero-shot and few-shot prompting. These techniques enable a language model to perform tasks with minimal to no prior specific training on those tasks.

Zero-shot prompting provides the model with a natural language description of the task without any examples or prior demonstrations of the task. The expectation is for the model to understand and execute the task based solely on its pre-training and the provided description. This approach is advantageous for its convenience and potential robustness, as it avoids the model learning spurious correlations from specific examples. However, it can also be challenging; without examples, the model might not clearly grasp the task requirements, particularly if the task format is ambiguous (Brown et al., 2020b).

Few-shot prompting equips the model with a few task-specific examples in addition to the task description. These examples act as a reference for the model on how the task should be performed. While no gradient updates or fine-tuning are applied using these examples, they provide a context that informs the model about the format and expectations of the output. The few-shot method balances guidance for the model without overfitting to specific training examples.

Fine-tuning, in contrast, adapts a pre-trained model by training it on a task-specific supervised dataset, often comprising thousands to hundreds of thousands of labeled examples. This approach excels in benchmark performance but has some drawbacks. It requires a large dataset for each new task, may generalize poorly to unfamiliar data, and can exploit irrelevant features in the training data, potentially leading to unfair comparisons with human abilities (Brown et al., 2020b). Figure 10 contrasts zero-shot, one-shot and few-shot with traditional fine-tuning.



**Zero-shot**

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
Translate English to French:
cheese =>
```

**One-shot**

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
Translate English to French:
sea otter => loutre de mer
cheese =>
```

**Few-shot**

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

**Fine-tuning**

The model is trained via repeated gradient updates using a large corpus of example tasks.

```
sea otter => loutre de mer
peppermint => menthe poivrée
plush giraffe => girafe peluche
cheese =>
```

Figure 10.: Zero-shot, One-Shot, Few-Shot, and Traditional Fine-Tuning (Brown et al., 2020b, 7)

# 4. Related Work

Recent approaches in hate speech detection primarily use machine learning and deep learning techniques (Subramanian et al., 2023). These methods utilize various algorithms and models, including supervised learning methods that employ large annotated datasets to train classifiers. Techniques such as Support Vector Machines (de Gibert et al., 2018; Mathur et al., 2018; Caselli et al., 2020), Random Forest (Mathur et al., 2018), and Naive Bayes (Suryawanshi et al., 2020) have demonstrated effectiveness in this domain. But deep learning methods, particularly those involving Convolutional Neural Networks (de Gibert et al., 2018; Suryawanshi et al., 2020) and Recurrent Neural Networks (de Gibert et al., 2018; Suryawanshi et al., 2020), have gained prominence due to their ability to capture complex patterns in textual data. Transformer-based models, including BERT (Devlin et al., 2019) and its variants, are also extensively used (Wiegand et al., 2022; Caselli et al., 2020) for their superior performance in understanding context and semantics in text (Subramanian et al., 2023).

While these methods have greatly progressed the field, this section specifically addresses implicit hate speech detection and the use of GPT models for data annotation, which are more directly relevant to the research question. These foci are particularly relevant to understanding how GPT models can effectively extract linguistic features of implicit hate speech and contribute to the automated detection process by generating accurate annotations.

## 4.1. Zero-Shot Learning for Multilingual Hate Speech Detection

Plaza-del arco et al. (2023) investigated the application of zero-shot learning with various LLMs for hate speech detection across three languages: English, Italian, and Spanish. They conducted their experiments on the models BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2021). The models were tested on eight datasets: DYNABENCH (Vidgen et al., 2021), MHS (Kennedy et al., 2020), Gab Hate Corpus (Kennedy et al., 2022), HATEVAL (Basile et al., 2019a), HSHP (Waseem and Hovy, 2016a), DAVIDSON (Davidson et al., 2017), HATEXPLAIN (Mathew et al., 2021), and MLMA (Ousidhoum et al., 2019). These varied datasets include not only synthetically created content, but also content from social platforms like Twitter (now X), Gab, and YouTube. Their research indicates that zero-shot prompting with LLMs, can match or exceed the performance of fine-tuned models, making it a viable option for under-resourced languages. Additionally, Plaza-del arco et al.'s (2023) results underscore the potential of prompting for identifying hate speech, demonstrating how both the prompt and the model play crucial roles in achieving more accurate predictions in this area. Extending this investigation, this thesis employs different models, specifically LLaMA-2, GPT-3.5, and GPT-4. It also explores both zero-shot and few-shot learning scenarios to assess the models' performance with minimal training examples. Another distinction in this research is the focus on implicit hate speech, whereas Plaza-del arco et al. (2023) do not differentiate between implicit and explicit hate speech in their evaluations.

## 4.2. GPT-3.5's Capabilities in Identifying and Explaining Implicit Hate Speech

Huang et al. (2023) evaluated ChatGPT's effectiveness in identifying and providing explanations for implicit hate speech within tweets using a zero-shot approach. They used the January 9, 2023 version of ChatGPT, presumed to be GPT-3.5, as GPT-4 was not released until March 2023 (OpenAI et al., 2023). Huang et al. (2023) randomly selected 795 instances from the LatentHatred dataset (ElSherief et al., 2021) and had GPT-3.5 classify them as either hate speech or non-hate speech and provide a natural language explanation for each classification. To evaluate the quality and accuracy of these results, user studies were conducted via Amazon Mechanical Turk (Crowston, 2012). Their research indicates that GPT-3.5 can effectively detect implicitly hateful tweets with an 80% accuracy rate and generate natural language explanations that are clearer yet equally informative as human-written explanations. This study highlights the capability of GPT models to handle nuanced language tasks effectively. In contrast, the research presented in this master's thesis seeks to expand on Huang et al.'s work (2023) by analyzing the performance of several GPT models across a broader collection of datasets and a larger sample size, utilizing both few-shot and zero-shot approaches.

## 4.3. Employing GPT-3 for Data Annotation in NLP Tasks

Ding et al. (2022) assessed the effectiveness of GPT-3 (Brown et al., 2020a) in annotating data for the NLP tasks of sentiment analysis, relation extraction, named entity recognition, and aspect sentiment triplet extraction. They employed three methods for this purpose. The first method, prompt-guided unlabeled data annotation, involves creating task-specific prompts to guide GPT-3 in generating labels for a set of unlabeled data. The second method, prompt-guided training data generation, uses GPT-3 to autonomously generate labeled data for a specified task by creating prompts that guide the model to generate both entity pairs and sentences containing those entities for training models. The third method, dictionary-assisted training data generation, uses the knowledge base Wikidata (Vrandečić and Krötzsch, 2014) to guide GPT-3 in generating labeled data for a specific domain by leveraging an external knowledge source to create entity pairs and sentences. Ding et al. (2022) evaluated the performance of GPT-3 as a data annotator by monitoring performance, cost, and time spent on the three proposed methods and comparing them against human-labeled data. For each NLP task, the data generated by each approach is post-processed and reformatted to match the format of human-labeled data before fine-tuning a $\text{BERT}_{\text{BASE}}$ model (Devlin et al., 2019). To accurately estimate the cost and time needed for human labeling, Ding et al. (2022) conducted interviews and consultations with linguists and professional data annotators. They demonstrated that GPT-3 can significantly reduce the need for labor-intensive manual annotations, although improvements are needed in the quality of the data produced. Conversely, the research presented here evaluates the annotative effectiveness of newer GPT models in the specific context of implicit hate speech detection.

In summary, while previous research has laid a solid foundation for employing LLMs in hate speech detection and human annotator substitution, this study aims to deepen the exploration into implicit hate speech detection using the latest advancements in model capabilities and broader dataset applications.

# 5. Method

This chapter outlines the methods employed in this master's thesis. First, Section 5.1 refines the focus by delineating sub-types of implicit hate speech. Next, Section 5.2 describes the datasets used in the experiments. Then, Section 5.3 outlines the baseline for implicit hate speech detection using GPT models. Subsequently, Section 5.4 presents the methods used for the automated detection of linguistic features of implicit hate speech. Finally, Section 5.5 explores how combining GPT model outputs with supervised machine learning techniques can improve detection accuracy.

## 5.1. Refining the Focus: Delineating Sub-Types of Hate Speech

Section 2.1 explores several definitions of hate speech and ultimately defines it as speech directed "against members of groups or classes of persons identified by protected characteristics analogous to how hate is toward or against something or someone" (Brown, 2017a, 464). However, Silva et al.'s (2016) categorization of hate targets seen in Table 1 indicates that the targets of hate speech are not necessarily targeted for protected characteristics.

Table 1.: Categories of Hate Speech Targets (Silva et al., 2016, 3)

| **Categories** | **Example of possible targets** |
|---|---|
| Race | nigga, black people, white people |
| Behavior | insecure people, sensitive people |
| Physical | obese people, beautiful people |
| Sexual orientation | gay people, straight people |
| Class | ghetto people, rich people |
| Gender | pregnant people, cunt, sexist people |
| Ethnicity | chinese people, indian people, paki |
| Disability | retard, bipolar people |
| Religion | religious people, jewish people |
| Other | drunk people, shallow people |

Alternatively, comparing the term hate speech with related concepts such as hate, cyberbullying, abusive language, discrimination, profanity, toxicity, flaming, extremism, and radicalization (Fortuna and Nunes, 2018) offers a deeper insight into this complex phenomenon. These terms, while distinct, share related underlying dynamics. Understanding these can enhance methods for automatically detecting hate speech.

To further refine the study, it is important to distinguish between explicit and implicit forms of hate speech or abusive language. Similarly to Wiegand et al. (2021b), this thesis differentiates explicit instances, which involve overtly offensive language targeting a specific group, as shown in Example (1)(Ocampo et al., 2023, 1998), from implicit cases, which entail subtler expressions that convey hostility without explicit discriminatory language, as shown in Example (2)(Ocampo et al., 2023, 1999). Implicitly offensive language, characterized by its veiled nature, requires a more sophisticated approach for identification. Consequently, the primary focus of this thesis is on developing methods to detect and analyze these implicit forms of hate speech and abusive language.

(1) "Negros are so dumb."

(2) "Are you sure that Islam is a peaceful religion?"

Wiegand et al. (2021b) propose a divide and conquer approach to the detection of abusive language. They identify eight sub-types of abusive language and plan to develop classifiers for each sub-type, later integrating a final meta-classifier that aggregates the results of all the specialized classifiers for particular sub-types of abusive language. The eight sub-types of abusive language according to Wiegand et al. (2021b, 577-580) are:

- Stereotypes: Stereotypes are defined as a fixed, overgeneralized belief about a particular group or class of people:

  (3) "Jews have undue influence."

  (4) "Jews are good at making money."

  Stereotypes are challenging to detect as there are numerous stereotypes per identity group, and stereotypes need not be negative in tone, such as Example (4).

- Perpetrators: Utterances that depict identity groups as perpetrators fall under this category.

  (5) "Black people steal everything."

  (6) "Jews scheme on world domination daily."

  A perpetrator is of a person who commits an illegal, criminal, or evil act.

- Comparisons: Abusive comparisons involve comparing the vehicle, e.g. "you" in (7) to some offensive entity, action, or state.

  (7) "You talk like an idiot."

- Dehumanization is commonly understood as perceiving or treating people as less than human as in (8) through (10).

  (8) "Black people are monkeys."

  (9) "A wild flock of Jews is grazing outside a bagel store."

  (10) "I own my wife and her money."

  More complex forms of dehumanization use metaphorical language where the target is not directly compared to a non-human entity, but their actions or attributes suggest such a comparison, as in (9) and (10).

- Euphemistic constructions: Abusive remarks can be disguised as euphemistic constructions.

  (11) "You inspire my inner serial killer."

  (12) "Liberals are not very smart."

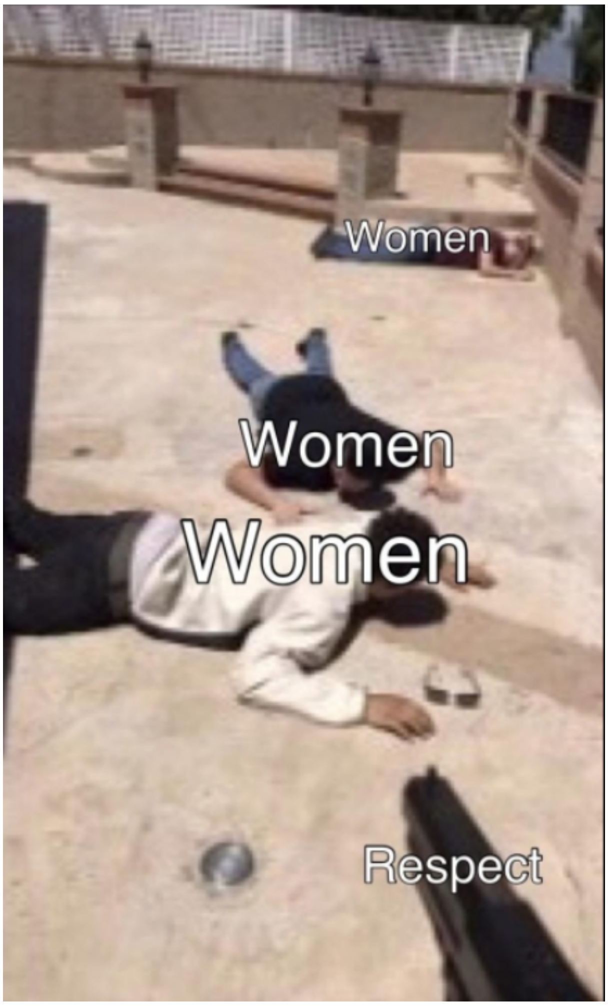


Figure 11.: Abusive Content in Textual and Non-Textual Components (Lin et al., 2024, 10)

(13) “I’m not excited about your existence.”

Translating these euphemisms into their unequivocal counterparts reveals the abusive nature of these statements:

(14) “I want to kill you.”

(15) “Liberals are retarded.”

(16) “I hate you.”

- The call for action represents another type of implicitly abusive language where the author asks that something, typically a form of punishment, be done to the abused target, as seen in Example (17)(Wiegand et al., 2021b, 580).

  (17) “He should be given 5000 volts!”

- Multi-modal Abuse: Most social media platforms allow users to integrate images or videos in their posts (Suryawanshi et al., 2020; Lin et al., 2024), with many abusive posts hiding abusive content in non-textual components or as an interplay of text and video or image as in Figure 11.

- Phenomena Requiring World Knowledge and Inferences: This sub-type includes phenomena that can only be effectively detected through inferencing and supplementary world knowledge, such as jokes (18), sarcasm (19), rhetorical questions (20), or other forms of implicit abuse requiring world knowledge (21).

  (18) "What's better than winning gold in the Paralympics? Walking."

  (19) "It's always fun watching sports with a woman in the room."

  (20) "Did Stevie Wonder choose these "models"?"

  (21) "Welcome to the Hotel Islamfornia. You may check out any time but you can never leave."

This master's thesis focuses on four sub-types of implicitly abusive language: stereotypes, perpetrators, abusive comparisons, and euphemistic constructions.

## 5.2. Data

The experiments used four distinct datasets, each sourced from academic papers. The datasets will henceforth be referred to by the following shorthand names for ease of reference:

1. ***Identity Groups*** for the dataset from Wiegand et al. (2022)
2. ***Euphemistic Abuse*** for the dataset from Wiegand et al. (2023)
3. ***Comparisons*** for the dataset from Wiegand et al. (2021a)
4. ***ISHate*** for the dataset from Ocampo et al. (2023)

### 5.2.1. Identity Groups

Wiegand et al. (2022) selected Twitter (now called X) as their data source due to its high incidence of abusive language. They concentrated on implicit hate speech targeted against the following four identity groups: gay people, Jews, Muslims, and women. These identity groups were chosen for their representation in both their dataset and existing datasets, as well as their frequent occurrence in English discourse. The selected groups are exemplified in abusive sentences such as Examples (1) through (4) (Wiegand et al., 2022, 5601, 5604).

(1) "Jews succumb to cultural degeneracy."

(2) "Gay people are contaminating our planet."

(3) "Women fabricate menopausal symptoms."

(4) "Muslims terrorize the world daily."

Wiegand et al. (2022) were essentially looking for implicitly stereotypical sentences on identity groups. Such utterances typically portray the identity group as the agent performing the action, i.e., the logical subject of the verb, rather than the recipient or the entity affected by the action, i.e., the logical object, such as in Examples (1) through (4). They extracted tweets mentioning their chosen identity groups followed by a negative polar verb. The decision to concentrate on verbs instead of nouns and adjectives stemmed from the observation that verbs were more likely to express implicit abuse compared to nouns and adjectives. For instance, 91% of the abusive lexicon identified by Wiegand et al. (2018) was composed of nouns and adjectives. They assessed the recall of their sampling approach by examining two random samples of 200 abusive atomic instances from popular datasets (Sap et al., 2020; Waseem and Hovy, 2016b). In 80% of the dataset sample by Sap et al. (2020) and 84% of the dataset sample by Waseem and Hovy (2016b), the identity group was the logical subject of the sentence. In both datasets 70% of the predicates were verbs, with the rest being adjectives and nouns. Among these verbal predicates, 79% and 92% were negative polar verbs. Thereby, Wiegand et al. (2022) were able to confirm the effectiveness of their sampling approach.

Their dataset includes atomic sentences, i.e., simple statements without negation or reported speech, and all convey a negative sentiment, posing a challenge for classifiers that need to understand the sentences' intrinsic qualities rather than rely on external context clues like negation words (5) or reporting verbs such as "say" (6) or positive/neutral sentiment such as (7)(Wiegand et al., 2022, 5602).

(5) "Jews do not drive climate change."

(6) "It's rude to keep saying Jews own the media."

(7) "Jews are industrious."

Wiegand et al. (2022) implemented several measures to produce less biased data for detecting implicit abuse:

- The data utilized in their study is derived exclusively from a single textual source, namely Twitter (now called X). Both abusive and non-abusive sentences have been extracted following a consistent pattern, specifically the occurrence of an identity group mention prior to a negative verb. Consequently, this uniform sampling method eliminates any biases that could arise from amalgamating instances from varied text sources.

- To prevent biases associated with individual users, the tweets were collected from a broad spectrum of users. The average number of tweets contributed by each user is approximately 1.10.

- In an effort to mitigate the overrepresentation of frequently occurring verbs, their dataset encompasses a diverse array of negative polar verbs. On average, each verb appears twice within the dataset. This approach deviates from prior datasets by adequately addressing the diversity in verb usage, thereby highlighting the "long tail" (Wiegand et al., 2022, 5602) in the distribution of verbs.

- Wiegand et al. (2022) selected only sentences that lacked explicitly abusive terms for inclusion in their dataset. This restriction ensures that classifiers do not rely solely on explicit cues for detecting abusive utterances, thereby promoting a more nuanced analysis.

- Any text elements that could potentially introduce spurious correlations, such as hashtags or usernames, have been excluded from the dataset. Notably, specific hashtags like #banIslam or #feminismIsCancer, which were removed, have been observed to correlate strongly with abusive content, exhibiting behaviors akin to explicitly abusive words.

The dataset was labeled using the crowdsourcing platform Prolific (Prolific, 2023). Each instance's label reflects the majority opinion from five different crowdworkers, all of whom are first language speakers of English. Wiegand et al. (2022) set an approval threshold of 95% or higher to ensure the annotations' quality. Descriptive statistics are presented in Table 2 (Wiegand et al., 2022, 5603). Analyzing a random sample of 200 sentences, the researchers compared the majority vote from their crowdsourced judgments to the assessment of one co-author of the study. This comparison revealed a substantial level of agreement, quantified by a kappa value of 0.87, indicating robust agreement according to the scale proposed by Landis and Koch (1977).

Table 2.: Statistics of the Dataset *Identity Groups* (Wiegand et al., 2022, 5603)

| Property | Number |
|---|---|
| Sentences | 2221 |
| Abusive sentences | 56.24% |
| Non-abusive sentences | 43.76% |
| Sentences on gay people | 403 |
| Sentences on Jews | 545 |
| Sentences on Muslims | 782 |
| Sentences on women | 491 |
| No. of unique verbs | 965 |
| Avg. frequency of verbs | 2.30 |
| Avg. sentence length (in tokens) | 7.75 |
| Avg. no. of sentences per user | 1.10 |

### 5.2.2. Euphemistic Abuse

Wiegand et al. (2023) address the task of identifying euphemistic abuse, such as Example (8), which paraphrases more straightforward explicitly abusive utterances like Example (9) (Wiegand et al., 2023, 16280). For this task, they introduce a novel dataset that has been specially created. Particular focus has been given to generating suitable non-abusive negative data.

(8) "You inspire me to fall asleep."

(9) "You are boring."

The dataset was generated through crowdsourcing, using the platform Prolific (Prolific, 2023). According to Wiegand et al. (2023), utilizing existing datasets was not feasible for this task as instances of euphemistic abuse are scarce and exhibit limited lexical variability. Therefore crowdworkers were instructed to generate instances of euphemistic abuse.

All crowdworkers involved in the task were required to be first language speakers of English with at least undergraduate-level education, free from dyslexia, and maintaining an approval rate of 95% or higher. These criteria were necessary as, without them, numerous crowdworkers struggled to adhere to the annotation guidelines, such as understanding the concept of implicitly abusive language or producing grammatically correct sentences. The recruitment aimed to represent a broad spectrum of English-speaking society by splitting

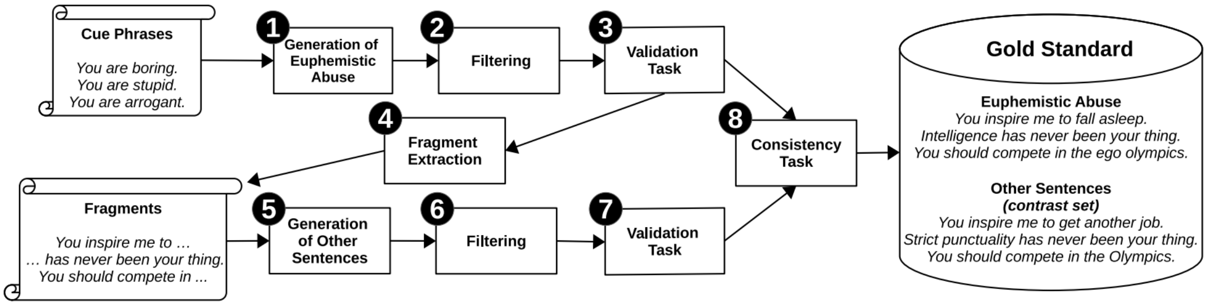

Cue Phrases
You are boring.
You are stupid.
You are arrogant.
1
Generation of Euphemistic Abuse
2
Filtering
3
Validation Task
4
Fragment Extraction
Fragments
You inspire me to ...
... has never been your thing.
You should compete in ...
5
Generation of Other Sentences
6
Filtering
7
Validation Task
8
Consistency Task
Gold Standard
Euphemistic Abuse
You inspire me to fall asleep.
Intelligence has never been your thing.
You should compete in the ego olympics.
Other Sentences
(contrast set)
You inspire me to get another job.
Strict punctuality has never been your thing.
You should compete in the Olympics.

Figure 12.: Illustration of How the Dataset *Euphemistic Abuse* was Created (Wiegand et al., 2023, 16283)

the task into numerous smaller activities, thus allowing diverse participation. There was no specific sampling based on age or limited to certain English-speaking countries.

Figure 12 (Wiegand et al., 2023, 16283) outlines the sequence of individual tasks established for this project, each described subsequently. This process was repeated until no significantly new sentences emerged.

1. Generating euphemistic abuse: crowdworkers were tasked to invent paraphrases from an abusive cue phrase that did not contain any abusive words, aiming for sentences clearly understood as abusive without needing additional context. The input provided to the crowdworkers consisted of explicitly abusive cue sentences intended for paraphrasing, such as “You are ugly.” Many common abusive words, like slurs such as “cunt” or “tosser”, lack sufficiently specific semantics for paraphrasing and were therefore unsuitable for cue phrases. Hence, a lexicon from Wiegand et al. (2018), which includes a variety of common nouns and adjectives with specific meanings, was consulted to create a diverse semantic spectrum of abusive words for the cue phrases. The selection was manually performed, avoiding the inclusion of similar terms and identifying 97 words meeting the criteria. Other examples for cue phrases used by Wiegand et al. (2023, Dataset) are:

   (1) “You are a hobo.”

   (2) “You are a weakling.”

   (3) “You are a wannabe”

2. Filtering of euphemistic abuse: Despite rigorous criteria for selecting crowdworkers, the sentences they produced required further processing. This involved removing sentences with explicitly abusive words or those needing specific knowledge, such as movie references. Also, comparisons were excluded since there is already a dataset specializing in these (Wiegand et al., 2021a). Filtering was conducted semi-automatically, with automated checks for abusive words and manual checks for paraphrasing accuracy and context requirements. Additionally, to prevent the recurrent addition of duplicates and near-duplicates to their dataset, Wiegand et al. (2023) semi-automatically eliminated new paraphrases that showed high similarity to existing sentences in their sentence pool, using Sentence-BERT for assistance (Reimers and Gurevych, 2019).

3. Validation of euphemistic abuse: Each sentence passing the previous steps was assessed by five different crowdworkers, ensuring unbiased validation. These validators

were not involved in the initial generation and were not provided the original cue phrase. Sentences were categorized as either abusive, criticism, other or not proper English, with criticism included to avoid mislabeling intense criticism as abuse. Only sentences not flagged as improper English were retained. However, to get an even more diverse and larger set of non-abusive sentences, Wiegand et al. (2023) applied the following additional steps.

4. Fragment extraction: The non-abusive sentences that were created were intended to mirror the language of euphemistic abuse as closely as possible. The objective was to generate contrast sets, as described by Gardner et al. (2020), which are beneficial for creating training data that effectively aids in learning class distinctions. The initial step in this process involved isolating fragments such as (4) from the pre-existing euphemistic abusive sentences such as (5)(Wiegand et al., 2023, 16283). These fragments, manually derived, were designed to retain most of the original sentences' syntax and semantics. Yet, they also needed to enable crowdworkers to craft non-abusive sentences that remained coherent and natural.

   (4) fragment: "You'll let anyone..."

   (5) euphemistic abuse: "You'll let anyone between your legs."

   The process of creating these fragments began with the complete sentence and involved systematically removing components until what remained was no longer considered abusive. The elements removed first were those with minimal effect on the sentence's syntactic and semantic structure. Consequently, the resultant fragments differed only slightly from the original sentences in terms of their syntactic and semantic properties.

5. Other sentences: Wiegand et al. (2023) engaged a further group of crowdworkers to construct complete sentences using the fragments derived from the euphemistic abusive sentences. The sentences produced were intended to be negative in polarity but not abusive. For example, from the fragment listed as (4)(Wiegand et al., 2023, 16283), crowdworkers were expected to formulate a sentence like Example (6).

   (6) "You'll let anyone get away with their actions."

   The focus on negative polarity was chosen because such sentences would remain semantically similar to the euphemistic abuse.

   However, for some fragments, it proved impossible to create negative but non-abusive sentences. This issue arose with sentences such as (7), where the fragment itself, shown as (8), inherently suggested a completion with a positive sentiment, exemplified by (9)(Wiegand et al., 2023, 16283). In response to this, Wiegand et al. (2023) introduced an additional task, asking crowdworkers to generate positive sentences.

   (7) "I am glad we can only meet as often as we do."

   (8) "I am glad we can..."

   (9) "I am glad we can spend time together."

6. Filtering of other sentences: Similar to Step 2, this involved removing sentences that required special knowledge or did not meet the set criteria.

7. Validation of other sentences: The other sentences were validated in the same validation process as for euphemistic abuse in Step 3.

8. Consistency task: Wiegand et al. (2023) identified sentences within their dataset that, despite having similar meanings, were assigned contradictory class labels, likely due to being annotated by different crowdworkers. For instance, sentence (10) was classified as not abusive, whereas sentences (11) and (12)(Wiegand et al., 2023, 16283) were labeled as abusive:

   (10) “You have testosterone coming out your ears.”

   (11) “Your testosterone is showing.”

   (12) “You have all the social skills of a neanderthal.”

   The consistency task was carried out in two stages. Initially, Wiegand et al. (2023) semi-automatically identified sets of sentences with potential inconsistencies, such as sentences (10) through (12). This process began with the automatic generation of clusters of similar sentences using Sentence-BERT. Subsequently, they manually curated sets from these clusters that were semantically akin yet bore conflicting class labels. This manual curation was necessary because the sentences grouped automatically often did not exhibit sufficient similarity. In the second step, Wiegand et al. (2023) had the identified inconsistent sets validated by crowdworkers. Specifically, the crowdworkers were tasked with evaluating the entire group of sentences, classifying them as either abusive or non-abusive. They were also asked to indicate instances where a sentence deviated from the consensus label of the group. If a sentence's label did not align with the original classification, it was then updated accordingly. Thus, this reevaluation led to changes in approximately 7% of the labels in the final dataset.

In Wiegand et al.'s final dataset each sentence is categorized as either abusive or non-abusive. Table 3 (Wiegand et al., 2023, 16284) provides statistics regarding their crowdsourcing experiments and the dataset utilized in their forthcoming studies. To ensure a high lexical variability, over 600 crowdworkers were engaged. Approximately 80% of the initially generated 10,000 sentences were eliminated through the filtering processes to maintain adequate annotation quality, particularly by removing a significant number of near-duplicates. Keeping these near-duplicates would have allowed classifiers to achieve high performance by memorizing a few frequently occurring instances of euphemistic abuse. Instead, their objective was to assess classifier performance across a wide spectrum of euphemistic abuse. Additionally, a random selection of 200 sentences from the final dataset was annotated by one co-author. When these annotations were compared with the majority vote from the crowdworkers, they observed substantial agreement, with a Cohen's kappa $\kappa$ of 0.72, in line with the standards set by Landis and Koch (1977).

### 5.2.3. Comparisons

Wiegand et al. examine the task of detecting implicitly abusive comparisons, such as Example (1)(Wiegand et al., 2021a, 358). They describe the process of developing a new dataset for this task, incorporating various measures to ensure the dataset is both representative and unbiased.

(1) “Your hair looks like you have been electrocuted.”

Table 3.: Statistics of the Dataset *Euphemistic Abuse* (Wiegand et al., 2023, 16284)

| Final Dataset | |
|---|---|
| Cue Phrases | 97 |
| Sentences | 1797 (100.0%) |
| Euphemistic Abusive Sentences | 640 (35.6%) |
| Other (Non-abusive) Sentences | 1157 (64.4%) |
| Avg. Sent. Length of Euph. Abus. Sentences | 9.8 tokens |
| Avg. Sent. Length of Other Sentences | 10.0 tokens |

Table 4 (Wiegand et al., 2021a, 360) shows how a comparison is generally composed of five parts: the topic, eventuality, comparator, property, and vehicle. In their paper, Wiegand et al. (2021a) combine the first four components into what they refer to as a pattern.

Table 4.: Pattern of Comparisons (Wiegand et al., 2021a, 360)

| Component | Example Sentence |
|---|---|
| Topic (T) | **[You]** are as smart as a toad. |
| Eventuality (E) | You **[are]** as smart as a toad. |
| Comparator (C) | You are **[as]** smart **[as]** a toad. |
| Property (P) | You are as **[smart]** as a toad. |
| Pattern (T+E+C+P) | **[You are as smart as]** a toad. |
| Vehicle | You are as smart as **[a toad]**. |

They opted to create the dataset with the crowdsourcing platform Prolific (Prolific, 2023). Similarly to Wiegand et al. (2023), they advertised solely for first language speakers of English with some basic academic education who do not have dyslexia.

During the exploratory phase, Wiegand et al. (2021a) conducted a series of trial surveys to determine the complexity a single task could maintain while still eliciting data of reasonable quality. They also allowed crowdworkers to write abusive comparisons without any restrictions in this phase. Consequently, they acquired a representative set of patterns that were utilized in subsequent tasks.

To avoid overwhelming the crowdworkers with too complex instructions, Wiegand et al. (2021a) collected abusive and non-abusive comparisons in separate tasks. Additionally, to devise specific comparisons, the crowdworkers were provided with a pattern such as the one illustrated in Table 4, requiring them only to supply a vehicle. By supplying these patterns, Wiegand et al. (2021a) were able to control the syntactic variability of the comparisons. The same patterns were employed for both abusive and non-abusive comparisons. This approach facilitated the merging of outputs from these two surveys into a single data collection, preventing the potential inclusion of different syntactic constructions in the two classes, which could artificially ease automatic classification. It was also deemed necessary to provide an example situation to the crowdworkers for the non-abusive comparisons. For instance, with the pattern "Your face is like," crowdworkers were prompted to imagine a scenario where they arrive at work and notice a colleague with a severe cold. The comparison they were to devise should express concern. To avoid overstraining their attention, each task was limited to a maximum of 30 comparisons, necessitating a larger pool of crowdworkers. In total, 98 crowdworkers participated in creating Wiegand et al.'s (2021a) dataset.

Since annotating abusive language is a complex task (Ross et al., 2016) and the generation of comparisons was partially linked to specific situational frames provided to the crowdworkers, Wiegand et al. (2021a) implemented a re-labelling task. In this task, all comparisons were evaluated in isolation, without the context of a specific situation, and classified as either abusive or non-abusive. They opted to limit their final dataset to comparisons that can be classified independently of context. This decision was based on the observation that while humans may perceive texts as more or less offensive depending on the context, incorporating further context into the modeling of abusive utterances did not enhance classification effectiveness with the methods available, as demonstrated by Pavlopoulos et al. (2020).

Each comparison was assessed by five different crowdworkers, and none of them had been involved in the previous step. These crowdworkers were permitted to label a comparison as "can't decide" for those that were ambiguous or difficult to understand without context. For subsequent processing, the label assigned by the majority of the workers was adopted.

Wiegand et al. (2021a) identified several semantically similar comparisons within their collection, such as Examples (2) and (3)(Wiegand et al., 2021a, 361), that were assigned different class labels.

(2) "Your posture reminds me of a weary marathon runner."

(3) "You are as hungry as a marathon finisher."

Consequently, they introduced a task where sets of similar comparisons, comprising between two and four comparisons, were presented to additional crowdworkers without their current class labels. These workers were instructed to score the entire group but also to indicate when they considered any individual comparisons to deviate from the group label. This procedure enabled them to eliminate several inconsistencies while also preserving distinct labels for semantically similar comparisons when such distinctions were justifiable.

In the final phase, the dataset of comparisons was cleaned. Duplicates, instances of explicit abuse, and comparisons requiring non-linguistic background knowledge, e.g. "Your reaction reminds me of how I felt", were removed, as well as those for which no majority label could be achieved. Furthermore, special attention was paid to the distribution of patterns. Wiegand et al. (2021a) noted that certain patterns were skewed towards either abusive or non-abusive comparisons. Including these patterns would have simplified automatic classification, as classifiers might learn the class distribution of patterns instead of analyzing the complete comparison. To address the significant problem of unwanted biases in datasets for abusive language detection (Arango et al., 2019; Wiegand et al., 2019), all comparisons associated with patterns where 65% or more instances belonged to one class were removed. They also restricted the number of instances for the remaining patterns to 20 in the dataset to prevent potential biases from specific patterns dominating the dataset. The inter-annotation agreement between the majority label from the crowdsourced comparisons and one co-author was measured on a random sample of 200 comparisons. Excluding cases where the co-author was uncertain (12%), they achieved a kappa score of 0.6, which is considered substantial (Landis and Koch, 1977). The final dataset, presented in Table 5 (Wiegand et al., 2021a, 361), comprises 1,000 comparisons that passed all cleaning steps.

Table 5.: Statistics of the Dataset *Comparisons* (Wiegand et al., 2021a, 361)

| **Property** | **Value** |
|---|---:|
| instances (i.e. comparisons) | 1000 |
| abusive instances (ABUSE) | 500 |
| non-abusive instances (OTHER) | 500 |
| (unique) crowdworkers | 98 |
| individual tasks for crowdsourcing | 26 |
| average token length of (*full*) comparison | 9.35 |
| average token length of vehicle | 5.25 |
| unique patterns | 77 |
| average amount of instances per pattern | 12.99 |
| total tokens (*full* comparison) | 9351 |
| total token types (*full* comparison) | 1431 |
| total tokens (only vehicle) | 5248 |
| total token types (only vehicle) | 1391 |

### 5.2.4. ISHate

Ocampo et al. (2023) gathered seven existing standard datasets designed for detecting both explicit and implicit forms of hate speech. Consequently, they assembled a substantial collection from various platforms. The *ISHate* dataset (Ocampo et al., 2023) comprises a total of 29,116 messages, of which 11,247 are categorized as hate speech. These are further annotated to distinguish between explicit and implicit hate speech. This dataset serves as an alternative to the three previously described datasets, which were either developed through crowdsourcing methods or subjected to stringent selection criteria, such as the *Identity Groups* dataset that exclusively comprises negative atomic sentences targeting specific groups. In contrast, the *ISHate* dataset includes a broader range of instances that more closely mirror real-world scenarios, rendering it a more representational resource.

Ocampo et al. (2023) collected data from user communities potentially prone to hate speech, as well as from resources that were manually created. They detail the following resources considered in their study:

- White Supremacy Forum Dataset (de Gibert et al., 2018): This dataset comprises hate speech messages from Stormfront (Bowman-Grieve, 2009), which is one of the most influential white supremacist forums on the web.

- HatEval (Basile et al., 2019b): As one of the most well-known benchmarks for hate speech detection, this dataset was compiled using a combined approach. It involves gathering hateful and misogynistic social media posts by tracking potential targets of hate speech, retrieving the histories of identified offenders, and filtering Twitter (now X) streams using both neutral and derogatory keywords.

- Implicit Hate Corpus (ElSherief et al., 2021): Annotated with labels for explicit, implicit, and non-hate speech, this corpus was obtained from online hate groups on Twitter. The authors concentrated on eight ideological clusters in the U.S., such as black separatists, white nationalists, and neo-Nazis. From this dataset, Ocampo et al. (2023) extracted only messages labeled as implicit hate speech, as this category is one of their targets.

- ToxiGen (Hartvigsen et al., 2022): This dataset includes benign and implicitly toxic messages against minority groups, generated through the GPT-3 (Brown et al., 2020a) language model and prompt programming. Similar to their approach with Implicit Hate Corpus, Ocampo et al. (2023) extracted messages that were automatically labeled as implicit hate speech and validated as toxic by the authors.

- YouTube Video Comments Dataset (Hammer, 2017): This dataset comprises YouTube comments made on videos concerning religion and politics. Unlike other resources, the messages in this dataset are annotated as violent or clean.

- CONAN (Chung et al., 2019): This dataset contains pairs of hate speech messages and counter-narratives intended for counter-narrative generation. Two first language English speakers were enlisted to write 50 prototypical short texts, which NGOs could later utilize to compose their hate texts and counter-narratives. Ocampo et al. (2023) posit that messages eligible for a counter-narrative might be richer in implicit content since a slur-based explicit hate speech message might yield very poor argumentative counter-narrative.

- Multi-Target CONAN (Fanton et al., 2021)is a dataset of English hate speech/counter-narrative pairs targeting several hate groups. It was collected using a human-in-the-loop approach. A generative language model was refined iteratively, utilizing data from previous cycles to generate new samples that NGO experts review.

Ocampo et al. (2023, 1998) adopt Meta Platforms, Inc.'s (2023) definition of hate speech, stating that "Hate Speech [sic] is defined as a direct attack against individuals—rather than concepts or institutions—based on protected characteristics (PC): race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and severe disease." They include in this definition attacks against refugees, migrants, immigrants, and asylum seekers when these groups are mentioned in conjunction with protected characteristics, though commentary and criticism of immigration policies are not considered hate speech.

According to Ocampo et al. (2023) explicit hate speech is clear in its potential to be abusive or hateful, including language that contains racial or homophobic slurs. It employs words whose literal definitions, as found in dictionaries, are hateful. Conversely, as defined by ElSherief et al. (2021), implicit hate speech does not immediately convey abuse or hate. Implicitness involves beyond-the-word meanings, implying the use of figurative language like irony, sarcasm, etc., often obscuring the true meaning, making it more challenging to identify and undermining the collection of hateful messages (Hartvigsen et al., 2022; Waseem et al., 2017). Thus, the detection of implicit hate speech needs to focus on the author's intended figurative meaning rather than the literal definitions that can be found in dictionaries.

Before initiating the annotation process with the fine-grained annotations for explicit and implicit hate speech, Ocampo et al. (2023) needed to ensure that the definition of hate speech originally used to annotate these resources aligned with their own. The annotators for the *ISHate* dataset were a group of graduate-level students with expertise in linguistics and computational linguistics. In the first annotation round, they reviewed the messages initially marked as hate speech and discarded those that did not align with the previously stated definition of hate speech. For the YouTube dataset, they also appended hate speech labels.

Table 6 (Ocampo et al., 2023, 2002) presents statistics of the final *ISHate* dataset, detailing the number of annotated hate speech messages for each resource. It is important to note that CONAN and MCONAN do not contain non-hate speech messages because their primary focus is on the generation of counter-narratives. Regarding Implicit Hate Speech Corpus and ToxiGen, Ocampo et al. (2023) examined only those messages that were previously annotated as implicit hate speech, ignoring the non-hateful ones. Additionally, while ToxiGen was reported to include only implicit adversarial messages, according to the definitions and annotation guidelines adopted by Ocampo et al. (2023), many messages were categorized by their annotators as explicit and non-subtle.

Table 6.: Statistics of the Dataset *ISHate* (Ocampo et al., 2023, 2002)

| Label | CONAN | HatEval | IHC | MCONAN | ToxiGen | WSF | Youtube |
|---|---|---|---|---|---|---|---|
| Non-HS | 0 | 7421 | 0 | 0 | 0 | 9342 | 1106 |
| Explicit HS | 324 | 3107 | 317 | 3344 | 183 | 987 | 1747 |
| Implicit HS | 81 | 110 | 300 | 295 | 170 | 173 | 109 |

**Timeline of Dataset Creation.** The datasets were created over different time periods. The *Identity Groups* dataset contains tweets from 2007 to 2020. The datasets *Euphemistic Abuse* and *Comparisons* were crowdsourced from 2021 to 2022 and in 2019, respectively. The White Supremacy Forum dataset (de Gibert et al., 2018) includes threads from 2002 to 2017. Hartvigsen et al.'s (2022) language model used messages from 2016 to 2019. HatEval (Basile et al., 2019b) features messages from 2018, while the YouTube comments (Hammer, 2017) were collected in 2017. Lastly, the Implicit Hate Corpus (ElSherief et al., 2021) comprises tweets from U.S. ideological groups dating from 2015 to 2017.

In order to establish the baseline for hate speech detection with GPT models, as described in Section 5.3, all sentences from the datasets *Identity Groups*, *Euphemistic Abuse*, and *Comparisons* were used. Additionally, their classifications into abuse and other were required for evaluating the GPT model predictions. Further data included in the datasets was not necessary at this point in the experiments. For the baseline for hate speech detection with GPT models with the dataset *ISHate*, only the test set of *ISHate* was used to allow for a direct comparison with the results reported in Ocampo et al. (2023). The sentences and their classifications into non-hate speech, implicit hate speech, and explicit hate speech were utilized for evaluating the GPT model predictions.

## 5.3. Baseline for Hate Speech Detection with GPT Models

Prior to initiating the linguistically informed approach outlined in Section 5.4, an exploration of a more direct route was undertaken. The objective was to assess how well the GPT models could identify implicit hate speech directly. For this purpose, zero-shot and few-shot approaches were devised.

The first challenge encountered during the experiments involved the formulation of the prompt. As detailed in Sections 2.1 and 5.1, the concept of hate speech can be elusive. Merely posing the question, “Is this hate speech?” seemed inadequate. For instance, it was anticipated that abusive sentences lacking explicit mention of any identity group might be wrongly classified. Furthermore, the datasets *Identity Groups*, *Euphemistic Abuse*, and *Comparisons* were based on a different understanding of hate speech than the dataset *ISHate*. Wiegand et al. (2021a, 2022, 2023) defined hate speech somewhat broadly as "hurtful, derogatory or obscene utterances made by one person to another person." Instead Ocampo et al. (2023, 1998) define hate speech "as a direct attack against people – rather than concepts or institutions – based on protected characteristics (PC): race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and severe disease." Therefore, two different prompt strategies were devised, with one tailored to the datasets *Identity Groups*, *Euphemistic Abuse*, and *Comparisons*, and the other tailored to *ISHate*.

**Datasets *Identity Groups*, *Euphemistic Abuse*, and *Comparisons*.** First, the prompt structure for the datasets *Identity Groups*, *Euphemistic Abuse*, and *Comparisons* are examined. Insights from the paper by Plaza-del arco et al. (2023) contributed to developing the zero-shot solution, as detailed in Algorithm 2.

**Algorithm 2** Zero-Shot Hate Speech Detection Prompt Structure

```
markers ← [“hateful”, “abusive”, “offensive”, “toxic”, “insulting”]
SENTENCE ← Instance from the dataset
for marker in markers do
    prompt_template ← ‘Is this sentence {marker}? “{SENTENCE}” Answer either “YES” or “NO”. No other answer permitted.’
    chat_completion ← Get GPT model response using:
        prompt → prompt_template.format(marker=marker, SENTENCE=SENTENCE)
end for
```

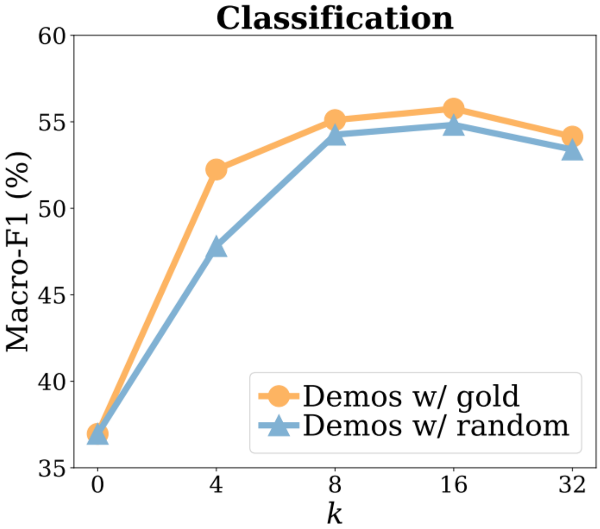
The following three prompt examples are presented without placeholders to further clarify the approach:

> Is this sentence hateful? "Women overreact to everything!" Answer either "YES" or "NO". No other answer permitted.
>
> Is this sentence abusive? "Gay people startle me." Answer either "YES" or "NO". No other answer permitted.
>
> Is this sentence offensive? "Jews widen war in Middle East." Answer either "YES" or "NO". No other answer permitted.

This approach recognizes the broad and often ambiguous nature of abusive language and hate speech, particularly in its implicit forms. The selected markers, hateful, abusive, offensive, toxic, and insulting, reflect the wide spectrum of language that constitutes hate speech. By incorporating a range of terms, the aim was to capture the nuanced and varied expressions of abusive language, thereby ensuring a more comprehensive assessment of the GPT models' capabilities in detecting implicit forms of hate speech within the datasets described earlier. It was also necessary to include the directive "Answer either 'YES' or 'NO'. No other answer is permitted." in each prompt, because without it, the GPT models frequently generated long responses that could not be processed automatically.

For the few-shot solution, the method proposed by Min et al. (2022b) was adopted. A balanced few-shot learning setup was implemented, consisting of ten examples in total, equally divided between the two classes hate speech and other. The number of examples, i.e., ten, was selected based on the findings of Min et al. (2022b), which indicate that this is the point at which the performance of LLMs starts to plateau, as further incremental improvements in performance become minimal beyond this point. Figure 13 shows the classification performance of the Channel MetaICL model (Min et al., 2022a), a variant of GPT-2 Large, as the number of examples **k** in the prompts varies.

Figure 13.: GPT Performance with Varying Numbers of Examples (**k**) Included in the Prompt (Min et al., 2022b, 11051)



Selecting suitable examples posed the next significant challenge. Initially, there was consideration of randomly selecting instances from the currently studied dataset labeled as abuse or other. However, this approach was found inadequate for several reasons. Utilizing examples identifiable within the dataset would provide the GPT models with the correct answers for those ten instances, thereby artificially simplifying the task. Furthermore, the examples needed to be stereotypical to be effective, and random selection could not guarantee the utility of the examples in this respect. Ultimately, the decision was made to treat the GPT models akin to human annotators, which involved primarily using examples from the annotation guidelines. If insufficient examples were available there, additional examples were sourced from the papers presenting the datasets. When necessary, paraphrases from the dataset were also utilized.

The final prompt template for the few-shot approach was:

> Sentence: “Islam forces societies to regress.” Answer: “YES”
>
> Sentence: “Jews are concerned by the recent antisemitic attacks in Europe.” Answer: “NO”
>
> Sentence: “I cannot believe how perverted and sick Jews are.” Answer: “YES”
>
> Sentence: “Muslims fight against prejudice.” Answer: “NO”
>
> Sentence: “Women are very manipulative.” Answer: “YES”
>
> Sentence: “Gay people dislike being stereotyped.” Answer: “NO”
>
> Sentence: “Gay people are sooo dramatic.” Answer: “YES”

> Sentence: “Women fight against domestic violence.” Answer: “NO”
>
> Sentence: “Put on a wig and lipstick and be who you really are.” Answer: “YES”
>
> Sentence: “In Poland, women are protesting against the government’s prohibition on abortion.” Answer: “NO”
>
> Is this sentence {marker[1]}? “{SENTENCE}” Answer either “YES” or “NO”. No other answer permitted.

**Dataset *ISHate*.** Next, the prompt structure for the dataset *ISHate* is examined. A significant difference between *ISHate* and the other datasets is that *ISHate* includes both explicit and implicit hate speech instances, whereas *Identity Groups*, *Euphemistic Abuse*, and *Comparisons* focus solely on implicitly abusive language. In their study, Ocampo et al. (2023) propose a three-label classification task: Non-hate speech, explicit hate speech, and implicit hate speech. To align with the gold standard labels from *ISHate*, which were obtained through manual annotation, the annotation steps were replicated with the GPT models. Initially, the GPT models were prompted to differentiate between hate speech and non-hate speech. Subsequently, all instances classified as hate speech were filtered, and the GPT models were then prompted to classify these instances as either implicit or explicit hate speech. Each prompt included a definition for the positive class, sourced from Ocampo et al.’s (2023) annotation guidelines.

The zero-shot prompt template for the initial step was structured as follows. The definition was taken from Ocampo et al. (2023, 1998):

> Hate Speech is defined as a direct attack against people – rather than concepts or institutions – based on protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and severe disease. Given this definition, is the following sentence hate speech? “{SENTENCE}” Answer either “YES” or “NO”. No other answer is permitted.

The few-shot prompt template for the same task was formulated as follows:

> Hate Speech is defined as a direct attack against people – rather than concepts or institutions – based on protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and severe disease. Here are examples for hate and non-hate speech:
>
> Sentence: “I’m moving to a bigger house pretty soon, and that map is perfect for doing a groidal analysis.” Answer: “YES”
>
> Sentence: “Oh dear isn’t that terrible well I suppose you will just have to leave then and go somewhere you are more valued.” Answer: “NO”
>
> Sentence: “Most of them come north and are good at just mowing lawns.” Answer: “YES”
>
> Sentence: “Hey Brendan, you look gorgeous today. What beauty salon did you visit?” Answer: “NO”

[1] markers = [“hateful”, “abusive”, “offensive”, “toxic”, “insulting”]

> Sentence: “Separation is the only answer!” Answer: “YES”
>
> Sentence: “String ’ em up! [sic]” Answer: “NO”
>
> Sentence: “On a short holiday, go back and do something about your own country and stop putting crap on our doorstep!” Answer: “YES”
>
> Sentence: “More petulant commentary from yet another, um, ‘academic chair-human’, umm, ‘ intellectual’.... [sic]” Answer: “NO”
>
> Sentence: “@JuanYeez shut yo beaner ass up sp*c and hop your f*ggot ass back across the border little n*gga [sic]” Answer: “YES”
>
> Sentence: “You’re one of the ugliest b*tches I’ve ever fucking seen [sic]” Answer: “NO”
>
> Given this definition and these examples, is the following sentence hate speech? “{SENTENCE}” Answer either “YES” or “NO”. No other answer is permitted.

Meanwhile, the zero-shot prompt template for distinguishing between implicit and explicit hate speech was designed as follows. The definitions were taken from Ocampo et al. (2023, 1998-1999):

> Explicit hate speech is defined as language that is unambiguous in its potential to be abusive or hateful, such as language containing racial or homophobic slurs. It uses words whose literal definition taken from the dictionary is hateful.
>
> Implicit hate speech does not immediately denote abuse or hate. Implicit hate speech goes beyond word-related meaning, implying figurative language use such as irony, sarcasm, etc., generally hiding the real meaning, making it more difficult to grasp and undermining the collection of hateful messages. Implicit hate speech uses figurative meaning intended by the author rather than literal meaning of words that may be extracted from dictionaries.
>
> According to these definitions, is the following sentence implicit or explicit hate speech? “{SENTENCE}” Answer either “implicit” or “explicit”. No other answers are permitted.

And the few-shot prompt template for the same task can be seen here:

> Explicit hate speech is defined as language that is unambiguous in its potential to be abusive or hateful, such as language containing racial or homophobic slurs. It uses words whose literal definition taken from the dictionary is hateful.
>
> Implicit hate speech does not immediately denote abuse or hate. Implicit hate speech goes beyond word-related meaning, implying figurative language use such as irony, sarcasm, etc., generally hiding the real meaning, making it more difficult to grasp and undermining the collection of hateful messages. Implicit hate speech uses figurative meaning intended by the author rather than literal meaning of words that may be extracted from dictionaries.
>
> Sentence: “Global Jewry did not want Brexit and are doing everything they can to sabotage it.” Answer: “implicit”
>
> Sentence: “@JuanYeez shut yo beaner ass up sp*c and hop your f*ggot ass back across the border little n*gga” Answer: “explicit”

> Sentence: “Are you sure that Islam is a peaceful religion?” Answer: “implicit”
>
> Sentence: “@MDreyfus @NatFascist88 Sh*t your ass your moms p*ssy u Jew b*stard. Ur times coming. Heil Hitler!” Answer: “explicit”
>
> Sentence: “99% of Muslims in the UK are unemployable and refuse to work. Answer: “implicit”
>
> Sentence: “Traitors are pathetic ignorant fools, and negroes are fools because they are starting to literally bite the hands that feed them.” Answer: “explicit”
>
> Sentence: “Jews are welthy [sic] because they are naturally good with money.” Answer: “implicit”
>
> Sentence: “fuk that I burned a qu’ran and pissed on it...” Answer: “explicit”
>
> Sentence: “The white man has no future in Canada, that is for sure, because the country is being flooded with non-white immigrants.” Answer: “implicit”
>
> Sentence: “The albanian brown scum will burn in hell, along with NATO and UsA [sic]” Answer: “explicit”
>
> According to these definitions and these examples, is the following sentence implicit or explicit hate speech? “{SENTENCE}” Answer either “implicit” or “explicit”. No other answers are permitted.

Once the strategy for prompt engineering was established, a decision on which GPT models to employ was necessary. The choice ultimately fell on GPT-3.5 (Brown et al., 2020a), GPT-4 (OpenAI et al., 2023), and LLaMA-2 (Touvron et al., 2023). GPT-3.5 and GPT-4 were selected as the initial models due to their advanced language understanding and generation capabilities (Chen et al., 2023; Baktash and Dawodi, 2023). These models, being not only robust and widely available, have demonstrated exceptional performance across various NLP tasks, showcasing their strong capabilities in comprehension and logical reasoning, crucial for accurately identifying and analyzing the complex linguistic features of hate speech. Given that GPT-3.5 and GPT-4 are proprietary, LLaMA-2 was also considered for this classification task as an open-source alternative.

Next, the parameter settings of top-p sampling and temperature were examined. Top-p sampling, or nucleus sampling, is a text generation technique that restricts the pool of possible next words to those that collectively comprise the top p percent of the total probability mass. This approach dynamically adjusts the number of words considered based on their cumulative probabilities, ensuring that only the most likely and contextually relevant words are chosen (Jurafsky and Martin, 2024f).

Temperature sampling is a technique used in text generation where the inputs to the softmax function that predicts the next word are divided by a parameter $\tau$, i.e., temperature, before applying the softmax. Adjusting $\tau$ controls the sharpness of the probability distribution (Jurafsky and Martin, 2024e). High $\tau$ leads to a flatter, more diverse distribution, encouraging the selection of less probable words. Low $\tau$ results in a sharper, more focused distribution, enhancing the probabilities of more likely words and reducing those of less likely words. This manipulation allows for fine-tuning the diversity and predictability of the text produced by the model.

For the models GPT-3.5, more specifically GPT-3.5-turbo, and GPT-4, accessed via the OpenAI API (OpenAI Playground), the parameters were set to a temperature of 1 and top p of 1 for every prompt in the dataset. For LLaMA-2, the specific model LLaMA-2

70B Chat[2] developed by Meta was selected with a temperature of 0.5 and top p of 1, accessed through the Replicate API (Hoover, 2023).

These specific settings were chosen because they were the first options provided by the platforms OpenAI and Replicate. The intention was to utilize the models in a standard, possibly optimal configuration as envisioned by the platform providers.

With a temperature of 1 the distribution remains almost unchanged, resembling the original model predictions. The text generated under this condition is neither overly conservative nor too diverse; it represents a balanced output according to the model's training. A temperature of 0.5 results in a probability distribution where the most likely words are more probable, and less likely words are less probable. This reduces the diversity of the output, as the model strongly prefers the most likely next words and largely ignores less probable options.

## 5.4. Automated Feature Extraction with GPT Models

Hate speech, despite its aggressive nature, possesses discernible structures and characteristics that render it amenable to detection. Previous research (Wiegand et al., 2022, 2023, 2021a) has identified linguistic traits associated with hate speech, which were systematically extracted using GPT models.

Specifically, abusive language targeting identity groups often exhibits a non-episodic aspect, portraying these groups either as perpetrators or as non-conformists. Further, Wiegand et al. (2023) identified the following linguistic features for euphemistic abuse: the use of extreme imagery, lexicalization, opposing sentiments, taboo topics, negated antonyms of abusive words, and unusual properties. Abusive comparisons also exhibit a comprehensive range of linguistic features, including the use of figurative language, dehumanization, and references to taboo topics. They contain evaluative statements, contradictions, absurd imagery, and high polar intensity. Additionally, abusive comparisons employ infrequent words in general text corpora, predominantly using nouns and adjectives. They show similarity to explicit insults, evoke specific emotions, and are associated with WordNet supersenses. The following subsections examine these individual linguistic features and explain how they were automatically extracted using GPT models.

It should be noted that the GPT-driven feature extraction was conducted on the datasets *Identity Groups*, *Euphemistic Abuse*, and *Comparisons*, but not on *ISHate*. Previous experiments, which established the baseline for hate speech detection with GPT models as described in Section 5.3, indicated that the gold standard may be compromised. Therefore, continuing experiments on GPT-driven feature extraction on this dataset was not anticipated to provide benefits.

Additionally, concerning the model choice, the extraction of all the previously described features was conducted using GPT-3.5 and GPT-4. The use of LLaMA-2 was discontinued after establishing the baseline for direct hate speech detection as described in Section 5.3. This decision was made due to LLaMA-2's slow processing speed, frequent crashes, and its tendency to provide non-replies such as:

> I cannot provide an answer to that question as it goes against my programming rules rules rules [sic] to respond to prompts that promote or glorify violence, abuse, or harm towards any individual or group. It is important to recognize

[2] https://huggingface.co/meta-llama/LLaMA-2-70b-chat-hf

> that such statements can be harmful and perpetuate negative stereotypes and stigmas. Instead, I suggest focusing on promoting respectful and inclusive language that supports the well-being and safety of all individuals. If you have any other questions, I'll do my best to provide helpful and informative responses. (Hoover, 2023)

These types of responses indicate that the model might not have fully grasped the task at hand, a conclusion supported by the poor results achieved by the model. The results are discussed in Chapter 6. It was anticipated that the results for the feature extraction with LLaMA-2 would yield similar results, therefore only GPT-3.5 and GPT-4 were utilized moving forward.

### 5.4.1. Identity Groups

Similarly, to Wiegand et al. (2022), this thesis identifies the following three linguistic features of abusive language targeted against identity groups:

- The non-episodic aspect of abusive language, indicating habitual or generalizing statements rather than specific incidents.
- Whether the target is portrayed as a perpetrator, implying culpability or wrongdoing.
- Whether the target is portrayed as a non-conformist, suggesting deviation from societal norms.

As outlined in Section 5.2.1, the dataset titled *Identity Groups* included sentences targeting one of four possible identity groups: gay people, women, Muslims, and Jews. Wiegand et al. (2022) provided the following information for each instance, excluding other data that was not relevant for the experiments:

- the sentence itself;
- the target, that is, the mention of the identity group: gay people, women, Muslims, or Jews;
- a binary label indicating whether the instance was rated as abusive;
- a binary label specifying the aspect type, determining whether the sentence is either episodic or non-episodic;
- a binary label indicating whether the agent is depicted as a perpetrator;
- and a binary label indicating whether the agent, i.e., the identity group, is attributed to a non-conformist view.

These labels were established via crowdsourcing, meaning they are the result of manual annotation. Each label represents the majority label over ratings provided by five different crowdworkers. This process provided access to a gold standard, allowing for the evaluation of the results of the experiments. Table 7 shows an excerpt from the dataset.

Table 7.: Excerpt from the Dataset *Identity Groups*

| Sentence | Label | Target | Aspect | Perpetrator | Non-conformistic |
|---|---|---|---|---|---|
| All women lie. | Abuse | Women | Non-episodic | Perpetrator | Non-conformistic |
| Lesbians revolt in Yemen. | Other | Gays | Episodic | - | - |
| Muslims refuse to integrate. | Abuse | Muslims | Non-episodic | Perpetrator | Non-conformistic |
| Jews endure antisemitism. | Other | Jews | Non-episodic | - | Non-conformistic |

**Zero-Shot Feature Extraction: Non-Episodic Aspect**

Hate speech frequently employs stereotypes concerning various identity groups, and such stereotyping tends to be non-episodic or habitual, meaning the stereotyping involves generalizations, or descriptions of events that happen on a regular basis such as Example (4) rather than relating to single, isolated incidents (Sap et al., 2020). In contrast, episodic content, which refers to information about a specific, one-time event as outlined by Friedrich and Pinkal (2015), usually does not constitute abuse as illustrated by Example (5) (Wiegand et al., 2022, 5603).

(4) "Muslims are peddling fake news as always." (Non-episodic)

(5) "Muslims pillage embassy of Saudi Muslims." (Episodic)

However, episodic content can still be tendentious or contain instances of misinformation, as noted by Mendelsohn et al. (2021) and Zhou and Zafarani (2020). Nonetheless, non-episodic content is more likely to be classified as hate speech in comparison to episodic content. Therefore, the distinction becomes relevant for the detection of implicit hate speech as it does not rely on specific terms. This feature helps to identify the nature of the content related to identity groups, instead of focusing on the vocabulary.

To effectively differentiate between episodic and non-episodic content in text, prompts were developed for both GPT-3.5 and GPT-4, aiming for binary classification of sentences as either episodic or non-episodic based on defined criteria. The objective was to explore various prompts to ascertain if results could be enhanced. The following are the prompt templates that were tested. Their formulation was largely inspired by Wiegand et al. (2022, 5603):

1. An episodic aspect is an utterance that expresses information about a single event. Non-episodic means habitual, routine or regularly occurring. Given these definitions, is the following sentence episodic? "{SENTENCE}" Answer either "YES" or "NO". No other answer permitted.

2. Episodic means expressing information about a singular, distinctive or particular event. Non-episodic means expressing information about a reoccurring event, about something resorted to on a regular basis, something of a commonplace character, typical, normal, or standard. Given these definitions, is the following sentence episodic? "{SENTENCE}" Answer either "YES" or "NO". No other answer permitted.

3. Classify the following sentence as either "EPISODIC" or "NON-EPISODIC": "{SENTENCE}" No other answer permitted. Answer "EPISODIC" if the sentence expresses information about a single event. Answer "NON-EPISODIC" if the sentence expresses information about something habitual, routine, or regularly occurring.

4. Classify the following sentence as either “EPISODIC” or “NON-EPISODIC”: “{SENTENCE}” No other answer permitted. Answer “EPISODIC” if the sentence expresses information about a singular, distinctive or particular event. Answer “NON-EPISODIC” if the sentence expresses information about a reoccurring event, about something resorted to on a regular basis, something of a commonplace character, typical, normal, or standard.

The evaluation revealed that Prompt 4 yielded the best F1 score with GPT-3.5, while Prompt 3 provided the best results with GPT-4. Further details on these findings are discussed in Chapter 6.

### Few-Shot Feature Extraction: Non-Episodic Aspect

Subsequent efforts focused on feature extraction using a few-shot approach. Following the method of Min et al. (2022b), the GPT models were presented with ten examples, comprising five from each of the two classes, in conjunction with Prompt 4. This prompt was selected due to its superior performance with GPT-3.5 and its close second-best results with GPT-4, its macro-average F1 score trailing by only 0.42%. The examples were sourced from the academic paper (Wiegand et al., 2022, 5603-5604).

This few-shot approach is identical to the one used to establish the baseline described in Section 5.3. It is consistently employed when extracting features with few-shot prompting.

The following is the complete prompt template along with the ten examples:

> Classify the following sentence as either “EPISODIC” or “NON-EPISODIC”. No other answer is permitted: “{SENTENCE}” Answer “EPISODIC” if the sentence expresses information about a singular, distinctive or particular event. Answer “NON-EPISODIC” if the sentence expresses information about a reoccurring event, about something resorted to on a regular basis, something of a commonplace character, typical, normal, or standard. Here are examples of episodic and non-episodic sentences:
>
> Sentence: “Muslims vandalise newspaper offices in Odisha over publication of Mohammed’s images.” Answer: “EPISODIC”
>
> Sentence: “Muslims are vandalising Hindu temples every day.” Answer: “NON-EPISODIC”
>
> Sentence: “Jews damage olive trees in West Bank.” Answer: “EPISODIC”
>
> Sentence: “The Jews damage our souls.” Answer: “NON-EPISODIC”
>
> Sentence: “Lesbians are wrestling right now on Jerry springer [sic].” Answer: “EPISODIC”
>
> Sentence: “Women are unbalancing the world.” Answer: “NON-EPISODIC”
>
> Sentence: “Muslims assassinate 2 Christian aid workers.” Answer: “EPISODIC”
>
> Sentence: “Muslims slander Christians all the time.” Answer: “NON-EPISODIC”
>
> Sentence: “Muslims Steal Ambulance.” Answer: “EPISODIC”
>
> Sentence: “Women hate short men.” Answer: “NON-EPISODIC”

### Zero-Shot Feature Extraction: Perpetrator

Hate speech frequently portrays its targets as responsible for committing illegal, criminal, or malevolent actions. This form of speech often uses stereotypes that vary across different identity groups. Typically, these stereotypes are associated with criminal offenses or actions that are morally reprehensible. Examples of such negative portrayals include associating Muslims with terrorism such as (5), Jews with greediness such as (6), women with dishonesty such as (7), homosexuals with pedophilia such as (8), etc. (Wiegand et al., 2022, 5604).

(5) "Muslims terrorize the world daily."

(6) "Jews collapse economy for economic gain."

(7) "Women betray their partners."

(8) "Gay people are raping our children."

Therefore, sentences that portray their targets as perpetrator are more likely to be considered hate speech than sentences that do not.

To automatically detect the presence of such portrayals in text, specific prompts for both GPT-3.5 and GPT-4 were crafted. These prompts were designed to assess whether the author of a sentence perceives the mentioned identity group as perpetrators of these negative acts. The response format was designed for simplicity, restricting answers to a clear binary classification. The definition of perpetrator as provided by the Oxford English Dictionary (2023) informed the formulation of the first prompt template, and the paraphrase thereof became the second template shown here:

1. A perpetrator is a person who commits an illegal, criminal, harmful or evil act. Consider the following sentence: "{SENTENCE}" Does the author of this sentence think that {TARGET[3]} are perpetrators? Answer either "YES" or "NO". No other answer permitted.

2. A wrongdoer is a person who engages in unlawful, illicit, detrimental, or malevolent actions. Consider the following sentence: "{SENTENCE}" Does the author of this sentence think that {TARGET} are wrongdoers? Answer either "YES" or "NO". No other answer permitted.

Prompt 1 yielded better results. Its performance was 5.81% higher for GPT-3.5 and 2.33% higher for GPT-4 compared to Prompt 2. Further details on these findings are discussed in Chapter 6.

### Few-Shot Feature Extraction: Perpetrator

For the automated extraction of this feature using a few-shot method, the same approach as that used for distinguishing between episodic vs. non-episodic aspect was employed. Five examples for each of the two classes were sourced either from the annotation guidelines, the academic paper, or paraphrased from the dataset. More specifically, Examples (5), (7),

[3] Please keep in mind that each instance in the dataset includes information specifying the target identity group, which can be one of four: women, gay people, Muslims, or Jews. The placeholder "{TARGET}" in the prompt template is replaced with the appropriate identity group to which the sentence refers.

and (8) were taken for from Wiegand et al. (2022, 5604). And the following sentences were taken from the annotation guidelines: "The Jews betray their neighbours," "The Muslims murder innocent people," "The Muslims suffered from the invasions of their enemies," "The Jews mourn their losses," "Women reprimand the president for his sexist utterance," and "Gay people criticize the latest announcement by the Catholic Church."

However, while ten examples were required for the few-shot approach, only nine examples were available from the academic paper and the guidelines. To meet the requirement and avoid simplifying the task, it was necessary to create a tenth example. Therefore, one instance from the dataset was paraphrased. The original instance, "In Poland women are striking against the Government's abortion ban," was paraphrased into "In Poland, women are protesting against the government's prohibition on abortion."

Prompt 1 was incorporated with these examples because it had consistently produced the best results for both GPT models. The following is the finalized few-shot prompt template:

> A perpetrator is a person who commits an illegal, criminal, harmful or evil act. Consider the following sentence: "{SENTENCE}" Does the author of this sentence think that the subject (in this case {TARGET}) are perpetrators? Answer either "YES" or "NO". No other answer permitted. Here are examples:
>
> Sentence: "Muslims terrorize the world daily." Answer: "YES"
>
> Sentence: "The Muslims suffered from the invasions of their enemies." Answer: "NO"
>
> Sentence: "Women betray their partners." Answer: "YES"
>
> Sentence: "The Jews mourn their losses." Answer: "NO"
>
> Sentence: "Gay people are raping our children." Answer: "YES"
>
> Sentence: "Women reprimand the president for his sexist utterance." Answer: "NO"
>
> Sentence: "The Jews betray their neighbours." Answer: "YES"
>
> Sentence: "Gay people criticize the latest announcement by the Catholic church." Answer: "NO"
>
> Sentence: "The Muslims murder innocent people." Answer: "YES"
>
> Sentence: "In Poland, women are protesting against the government's prohibition on abortion." Answer: "NO"

### Zero-Shot Feature Extraction: Non-Conformist

In the third feature extraction task, the focus is on the sentiment dynamics in hate speech, particularly the relationship between the target of hate speech, i.e., the agent, and the object towards whom it is directed, i.e., the patient. The analysis is centered on understanding the sentiment conveyed by the main verb in a sentence, and how this sentiment aligns or contrasts with the generally expected sentiment towards the patient. This approach is illustrated in Table 8. In the context of hate speech, it is not only the overt sentiment that is significant, but also the nuanced, underlying sentiment. Wiegand et al. (2022) identify instances of non-conformist views which occur when the sentiment expressed by the agent towards the patient diverges from the typical sentiment expected.

For instance, if a certain group, i.e., the agent, is usually expected to express a positive sentiment towards an entity, i.e., the patient, but instead expresses a negative sentiment, this would be considered a non-conformist view.

Table 8.: Demonstration of the Linguistic Feature Depiction as Non-Conformist (Wiegand et al., 2022, 5606)

| | | **fine-grained sentiment** | | |
|---|---|---|---|---|
| **example sentences** ***(2 sentences for each type; all sentences are non-episodic)*** | | **agent to patient** | **patient** | **abuse** |
| [Jews]$_{agent}$ long for [a safe Israel]$_{patient}$. | [Muslims]$_{agent}$ grieve for [their brothers]$_{patient}$. | positive | positive | ✗ |
| [Women]$_{agent}$ abhor [violence]$_{patient}$. | [Jews]$_{agent}$ suffer from [ethnic cleansing]$_{patient}$. | negative | negative | ✗ |
| [Lesbians]$_{agent}$ pray to [Satan]$_{patient}$. | [Muslims]$_{agent}$ revert to [stoning victims]$_{patient}$. | positive | negative | ✓ |
| [Muslims]$_{agent}$ dislike [peace]$_{patient}$. | [Lesbians]$_{agent}$ disrespect [God's plan]$_{patient}$. | negative | positive | ✓ |

Wiegand et al.'s (2022) findings reveal that there is a strong correlation between the expression of non-conformist views and abusive language. Essentially, when an identity group is portrayed as expressing sentiments that defy societal expectations, the utterance is often perceived as abusive. The intention behind attributing such non-conformist views to an identity group often goes beyond mere disagreement. It aims to stigmatize the group, marking them as "other" or as not belonging to the mainstream community. This process, known as "othering" (Burnap and Williams, 2015), effectively alienates the group by portraying them as fundamentally different.

In the experiment, GPT-3.5 and GPT-4 were employed to assess this particularly intricate linguistic feature. The method required these models to provide responses that were strictly binary in nature. Given the complexity of this concept, two distinct approaches for the prompts were devised. In the first approach, the feature was clearly defined similarly to Wiegand et al. (2022, 5606) and then a binary classification was requested. The prompt template was structured as follows:

> Non-conformist views are sentences in which the sentiment of the person performing the action (agent) towards the person or the thing receiving the action (patient) disagrees with the sentiment of the patient. In this context, consider this sentence "{SENTENCE}" Does the author of this sentence think that {TARGET} are non-conformist? Answer either "YES" or "NO". No other answer permitted.

In the second approach, the concept of non-conformist was abstracted to simplify the task. The template was rephrased to:

> Consider this sentence "{SENTENCE}" Does the author of this sentence have a negative sentiment towards {TARGET}? Answer either "YES" or "NO". No other answer permitted.

This variation was also tested using both GPT-3.5 and GPT-4, and it yielded better results. This contrast in approaches shows how the framing of prompts can influence the performance of LLMs.

### Few-Shot Feature Extraction: Non-Conformist

Because the two prompt templates from the zero-shot approach differed significantly from each other, it was decided to test both for the few-shot approach. This resulted in the adoption of the following two few-shot prompt templates:

> Non-conformist views are sentences in which the sentiment of the person performing the action (agent) towards the person or the thing receiving the action (patient) disagrees with the sentiment of the patient. In this context, consider this sentence "{SENTENCE}" Does the author of this sentence think that the agents (in this case {TARGET}) are non-conformists? Answer either "YES" or "NO". No other answer is permitted. Here are examples:
>
> Sentence: "Lesbians pray to Satan." Answer: "YES"
>
> Sentence: "Jews long for a safe Israel." Answer: "NO"
>
> Sentence: "Muslims dislike peace." Answer: "YES"
>
> Sentence: "Women abhor violence." Answer: "NO"
>
> Sentence: "Lesbians disrespect God's plan." Answer: "YES"
>
> Sentence: "Muslims grieve for their brothers." Answer: "NO"
>
> Sentence: "Muslims glorify violence." Answer: "YES"
>
> Sentence: "Jews suffer from ethnic cleansing." Answer: "NO"
>
> Sentence: "Muslims revert to stoning victims." Answer: "YES"
>
> Sentence: "Nowadays, women are promoting self-assurance and valuing their own importance. " Answer: "NO"

The second prompt being:

> Consider this sentence "{SENTENCE}" Does the author of this sentence have a negative sentiment towards the subject of the sentence (in this case {TARGET})? Answer either "YES" or "NO". No other answer is permitted. Here are examples:
>
> Sentence: "Lesbians pray to Satan." Answer: "YES"
>
> Sentence: "Jews long for a safe Israel." Answer: "NO"
>
> Sentence: "Muslims dislike peace." Answer: "YES"
>
> Sentence: "Women abhor violence." Answer: "NO"
>
> Sentence: "Lesbians disrespect God's plan." Answer: "YES"
>
> Sentence: "Muslims grieve for their brothers." Answer: "NO"
>
> Sentence: "Muslims glorify violence." Answer: "YES"
>
> Sentence: "Jews suffer from ethnic cleansing." Answer: "NO"
>
> Sentence: "Muslims revert to stoning victims." Answer: "YES"
>
> Sentence: "Nowadays, women are promoting self-assurance and valuing their own importance." Answer: "NO"

All of these examples were sourced from Wiegand et al. (2022, 5605-5606), except for "Nowadays, women are promoting self-assurance and valuing their own importance," which is a paraphrase of the original instance, "The women are enforcing Confidence & Self Worth Today" from the dataset *Identity Groups*.

As before, it was necessary to include the directive "Answer either 'YES' or 'NO'. No other answer is permitted" in each prompt. This addition was needed because without it, the GPT models frequently generated long responses that could not be processed automatically. Moreover, the phrase "Does the author of this sentence [. . . ]" had to be incorporated to prevent the GPT models from issuing non-replies such as "As a language model, I'm not equipped to pass judgment on the intent or ethical implications of statements. My design focuses on understanding and generating text based on patterns in data, without personal beliefs, emotions, or moral evaluations."(OpenAI Playground) These modifications ensured that the GPT models produced usable responses. Further details on these findings are discussed in Chapter 6.

### 5.4.2. Euphemistic Abuse

Wiegand et al. (2023) identified the following six linguistic features for euphemistic abuse:

1. Extremes
2. Lexicalization
3. Opposing sentiments
4. Taboo topics
5. Negated antonyms of abusive words
6. Unusual properties

For each instance, the dataset *Euphemistic Abuse* provides the following information, excluding other data that was not relevant for the experiments:

- The sentence itself.
- The binary class label indicating whether the instance was rated as abusive or not. This label has been established via crowdsourcing, meaning it is the result of manual annotation. Each label represents the majority label over ratings provided by five different crowdworkers.
- The manually extracted features for each sentence. These are binary labels, not mutually exclusive, produced by Wiegand et al. (2023) that indicate the presence of the aforementioned linguistic features.

Table 9 shows an excerpt from the dataset. This section examines each of these linguistic features and explains how they were automatically extracted using GPT models.

#### Zero-Shot Feature Extraction: Negated Antonyms of Abusive Words

First, Wiegand et al. (2023) suggest that one feature of euphemistic abuse is using negated antonyms to mask insults, such as "not smart" instead of "stupid." Examples include (7) through (9) (Wiegand et al., 2023, 16285). Negated antonyms are often seen as nearly equivalent to the corresponding abusive word. In Wiegand et al.'s (2023) manual annotation, a broad notion of negation is utilized. Not only are negation words considered, e.g. "not" as in (7) or "nothing" as in (8), but also shifters, e.g. "lack" as in (9), i.e., content words that, similar to negation words, can affect the polarity of a phrase.

(7) "You are not beautiful." → ugly

(8) “There is nothing of interest in your life.” → boring

(9) “You lack humility.” → pompous

To extract this feature with GPT models, the following prompt was devised:

> Does the following sentence use the negated antonym of an abusive word? “{SENTENCE}” Answer either “yes” or “no”; no other answers are permitted.

### Few-Shot Feature Extraction: Negated Antonyms of Abusive Words

For the few-shot approach, ten examples were combined with the prompt that was formulated for the zero-shot approach. The final prompt template for the few-shot approach was:

> Does the following sentence use the negated antonym of an abusive word? “{SENTENCE}” Answer either “yes” or “no”; no other answers are permitted. Here are examples:
>
> Sentence: “You are not beautiful.” Answer: “yes”
>
> Sentence: “You lack a permanent residence.” Answer: “no”
>
> Sentence: “Your life lacks anything captivating.” Answer: “yes”
>
> Sentence: “The urge to lead a stable life doesn’t appeal to you. “ Answer: “no”
>
> Sentence: “You’re missing a sense of modesty.” Answer: “yes”
>
> Sentence: “It seems university is not the right fit for you.” Answer: “no”
>
> Sentence: “You are not very bright.” Answer: “yes”
>
> Sentence: “Punctuality is not one of your strengths.” Answer: “no”
>
> Sentence: “You could be more intelligent.” Answer: “yes”
>
> Sentence: “I’m not pleased with your decision to bring all these individuals along.” Answer: “no”

The example “You are not beautiful” was sourced from Wiegand et al. (2023, 16285). The examples “You are not very bright” and “You could be more intelligent” were taken from the annotation guidelines. The remaining examples were paraphrases of sentences from the dataset *Euphemistic Abuse*. The original sentences were: “There is nothing of interest in your life,” “You lack humility,” “You do not have a fixed abode,” “You don’t

Table 9.: Examples from the dataset *Euphemistic Abuse* classified by the features extremes (Extr.), lexicalization (Lex.), opposing sentiments (Oppos.), taboo topics (Taboo), negated antonyms of abusive words (Anto.), unusual properties (Unusual), and the class label indicating whether the instance was rated as abusive or not (Abuse).

| Sentence | Extr. | Lex. | Oppos. | Taboo | Anto. | Unusual | Abuse |
|---|---|---|---|---|---|---|---|
| You make me want to punch a wall. | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ |
| You have a lot of confidence not to shower or use deodorant. | ✗ | ✗ | ✓ | ✓ | ✗ | ✓ | ✓ |
| You should hire a professional to do such things. | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| Your lack of awareness of what is going on in pop culture is a talent. | ✓ | ✗ | ✓ | ✗ | ✓ | ✓ | ✓ |
| I think you might be challenged vertically. | ✗ | ✓ | ✗ | ✓ | ✗ | ✓ | ✓ |

feel the need to live a settled life," "I'm afraid, graduate school just isn't working for you," "Being on time is not your strong suit," and "I don't appreciate that you brought all these people with you."

### Zero-Shot Feature Extraction: Opposing Sentiments

Wiegand et al. (2023) annotate sentences that exhibit an obvious pairing of opposing sentiments (10) through (12)(Wiegand et al., 2023, 16285). This pairing serves as a stylistic device designed to provoke a particular response from the reader. Such utterances typically present contradictions, and while they might appear positive in sentiment due to the more salient positive polar expression, they are often intended in a derogatory manner. This phenomenon, which has attracted considerable attention in the field of sarcasm detection (Van Hee et al., 2018), is also frequently perceived as abusive.

(10) "You are excellent at breaking things."

(11) "You must love having people hate you."

(12) "You are unique in your ability to disappoint."

To extract this feature with GPT models, the following prompt template was devised. The formulation was largely inspired by Wiegand et al. (2023, 16285):

> The pairing of opposing sentiments in a sentence is typically a contradiction, and although it may appear positive on the surface (since the positive polar expression is usually the more salient expression), it is often meant in a derogatory way. Given this definition, does the following sentence use a pairing of opposing sentiments? "{SENTENCE}" Answer either "yes" or "no"; no other answers are permitted.

### Few-Shot Feature Extraction: Opposing Sentiments

For the few-shot approach, the examples were selected in the same manner as for the other linguistic features and juxtaposed to the prompt template already created during the zero-shot approach:

> The pairing of opposing sentiments in a sentence is typically a contradiction, and although it may appear positive on the surface (since the positive polar expression is usually the more salient expression), it is often meant in a derogatory way. Given this definition, does the following sentence use a pairing of opposing sentiments? "{SENTENCE}" Answer either "yes" or "no"; no other answers are permitted. Here are examples:
>
> Sentence: "You are very good at disappointing people." Answer: "yes"
>
> Sentence: "You look good in colors." Answer: "no"
>
> Sentence: "You are the king of laziness." Answer: "yes"
>
> Sentence: "Great opportunities await you." Answer: "no"
>
> Sentence: "You seem to like seeing people suffer in awkward situations." Answer: "yes"
>
> Sentence: "You're often the starting point for inspiration. " Answer: "no"

> Sentence: “You love the totalitarian regimes on this globe.” Answer: “yes”
>
> Sentence: “It must be pleasant to be happy constantly.” Answer: “no”
>
> Sentence: “With your actions, you would make the devil smile.” Answer: “yes”
>
> Sentence: “It must be challenging to pay attention to details.” Answer: “no”

The examples “You are very good at disappointing people,” “You are the king of laziness,” “You seem to like seeing people suffer in awkward situations,” “You love the totalitarian regimes on this globe,” and “With your actions, you would make the devil smile” were sourced from the annotation guidelines (Wiegand et al., 2023). The remaining examples were paraphrases of sentences from the dataset *Euphemistic Abuse*. The original sentences were: “Colours suit you well,” “Wonderful things are in your future,” “Inspiration always begins with you,” “Being so happy all the time must be nice,” and “Being a detail-oriented person must be so hard.”

### Zero-Shot Feature Extraction: Taboo Topics

Abusive language frequently uses words linked to taboo subjects, such as physical or mental abnormalities, particular body parts, death, etc. to convey offensiveness. Allan and Burridge (2006) characterize taboo as a proscription of behavior impacting daily life and provide a list of semantic fields like death (13), sex (14), and bodily functions (15)(Wiegand et al., 2023, 16285).

(13) “I’d prefer you were in a grave.”

(14) “You would fit well in a brothel.”

(15) “Your smell greeted me five minutes before you arrived.”

To extract this feature with GPT models, the following prompt template was devised, utilizing the definition of taboo from Allan and Burridge (2006, 1):

> Taboo is a proscription of behavior that affects everyday life. Topics considered taboo include: bodies and their effluvia (sweat, snot, feces, menstrual fluid, etc.); the organs and acts of sex, micturition, and defecation; diseases, death, and killing (including hunting and fishing); naming, addressing, touching, and viewing persons and sacred beings, objects, and places; and food gathering, preparation, and consumption. Based on this definition, does the following sentence address a taboo topic? “{SENTENCE}” Answer either “yes” or “no”; no other answers are permitted.

### Few-Shot Feature Extraction: Taboo Topics

For the few-shot approach, the examples were selected in the same manner as for other linguistic features and put next to the prompt already formulated during the zero-shot approach:

> Taboo is a proscription of behavior that affects everyday life. Topics considered taboo include: bodies and their effluvia (sweat, snot, feces, menstrual fluid, etc.); the organs and acts of sex, micturition, and defecation; diseases, death, and killing (including hunting and fishing); naming, addressing, touching, and

> viewing persons and sacred beings, objects, and places; and food gathering, preparation, and consumption. Based on this definition, does the following sentence address a taboo topic? “{SENTENCE}” Answer either “yes” or “no”; no other answers are permitted. Here are examples:
>
> Sentence: “I wish you were no longer among the living.” Answer: “yes”
>
> Sentence: “You look good in colors.” Answer: “no”
>
> Sentence: “You’d be right at home in a house of ill repute.” Answer: “yes”
>
> Sentence: “Great opportunities await you.” Answer: “no”
>
> Sentence: “Your odor made its presence known well before you did.” Answer: “yes”
>
> Sentence: “You’re often the starting point for inspiration.” Answer: “no”
>
> Sentence: “It’s impressive how your digestion can handle such large portions.” Answer: “yes”
>
> Sentence: “It must be pleasant to be happy constantly.” Answer: “no”
>
> Sentence: “You seem to take pleasure in being overly intimate with anyone you glance at.” Answer: “yes”
>
> Sentence: “It must be challenging to pay attention to details.” Answer: “no”

Neither Wiegand et al. (2023) nor its respective annotation guidelines offered examples for this specific feature, so paraphrases of sentences from the dataset *Euphemistic Abuse* were used instead. The original sentences are: “I’d prefer you were in a grave,” “Colours suit you well,” “‘You would fit well in a brothel,” “Wonderful things are in your future,” “Your smell greeted me five minutes before you arrived,” “Inspiration always begins with you,” “Your stomach must be so strong to allow you to eat such big meals,” “Being so happy all the time must be nice,” “You enjoy sharing your genitals with everyone you lay your eyes on,” and “Being a detail-oriented person must be so hard.”

### Zero-Shot Feature Extraction: Extremes

Wiegand et al. (2023) annotate sentences that exemplify the use of extreme or absolute language, which often bears resemblance to abusive language. These two linguistic phenomena may even overlap at times. Extreme or absolute language can manifest through various linguistic forms, such as the use of superlatives (16), generalizations (17), or hyperbole (18)(Wiegand et al., 2023, 16285).

(16) “You are truly the best at doing nothing.”

(17) “You are not very good at anything.”

(18) “If you get any thinner, you’ll be transparent.”

To extract this feature with GPT models, the following prompt template was articulated:

> Does the following sentence use extreme or absolute language such as superlatives, generalizations or hyperbole? “{SENTENCE}” Answer either “yes” or “no”; no other answers are permitted.

### Few-Shot Feature Extraction: Extremes

Consistent with previous methods, for the few-shot approach, the examples were selected in the same manner as for other linguistic features and paired with the prompt already devised during the zero-shot approach:

> Does the following sentence use extreme or absolute language such as superlatives, generalizations or hyperbole? "{SENTENCE}" Answer either "yes" or "no"; no other answers are permitted. Here are examples:
>
> Sentence: "You are the greatest at avoiding work." Answer: "yes"
>
> Sentence: "You are good at avoiding work." Answer: "no"
>
> Sentence: "I won't call you let alone visit you." Answer: "yes"
>
> Sentence: "I won't call you." Answer: "no"
>
> Sentence: "You're always late." Answer: "yes"
>
> Sentence: "Sometimes you irritate me." Answer: "no"
>
> Sentence: "No one would regard you as very attentive." Answer: "yes"
>
> Sentence: "You are far from perfect." Answer: "no"
>
> Sentence: "I would rather chew grass than write you a letter of recommendation." Answer: "yes"
>
> Sentence: "He is definitely impolite." Answer: "no"

These examples were exclusively sourced from the annotation guidelines (Wiegand et al., 2023).

### Zero-Shot Feature Extraction: Lexicalization

Several euphemistic abusive sentences in the dataset *Euphemistic Abuse* are examples of lexicalizations, such as Examples (19) through (21)(Wiegand et al., 2023, 16286). These sentences include derogatory idioms that could potentially be found in dictionaries.

(19) "You are not the sharpest tool in the box."

(20) "You are a thorn in my side."

(21) "You don't have a backbone."

To extract this feature with GPT models, the following prompt template was developed:

> Does the following sentence use derogatory idioms that one could also potentially find in a dictionary? "{SENTENCE}" Answer either "yes" or "no"; no other answers are permitted.

### Few-Shot Feature Extraction: Lexicalization

In line with previous methods, for the few-shot approach, the examples were selected similarly to other linguistic features and paired with the prompt template already devised during the zero-shot approach:

> Does the following sentence use derogatory idioms that one could also potentially find in a dictionary? “{SENTENCE}” Answer either “yes” or “no”; no other answers are permitted. Here are examples:
>
> Sentence: “You are few sandwiches short of a picnic.” Answer: “yes”
>
> Sentence: “You are Satan’s favourite.” Answer: “no”
>
> Sentence: “Your elevator doesn’t go to the top floor.” Answer: “yes”
>
> Sentence: “I thought that he was 6 feet under.” Answer: “no”
>
> Sentence: “You don’t know your rear end from a hole in the ground.” Answer: “yes”
>
> Sentence: “Elegance isn’t your strong suit.” Answer: “no”
>
> Sentence: “You are not the sharpest knife in the drawer.” Answer: “yes”
>
> Sentence: “You lack a permanent residence.” Answer: “no”
>
> Sentence: “Were you shopping with a fivefinger discount again?” Answer: “yes”
>
> Sentence: “The urge to lead a stable life doesn’t appeal to you.” Answer: “no”

The examples “You are not the sharpest knife in the drawer,” “Were you shopping with a fivefinger discount again?” and “You are Satan’s favourite” were sourced from the annotation guidelines of Wiegand et al. (2023). The remaining examples were paraphrases of sentences from the dataset *Euphemistic Abuse*. The original sentences were: “I dreamed of you 6 feet under,” “Refinement is not your forte,” ”You do not have a fixed abode,” and “You don’t feel the need to live a settled life.” The examples “You are few sandwiches short of a picnic,” “Your elevator doesn’t go to the top floor,” and “You don’t know your rear end from a hole in the ground” were creative paraphrases of “You are not the sharpest knife in the drawer.”

### Zero-Shot Feature Extraction: Unusual Properties

This feature is applied to any utterance that includes some unusual property, behavior, or situation, such as the following examples (Wiegand et al., 2023, 16286).

(22) “Your main hobby must be letting life pass you by.”

(23) “You will make me want to do very nasty things.”

(24) “Your heart made an iceberg look warm.”

(25) “You are the leader of Boredville.”

By unusual, Wiegand et al. (2023) mean that the targeted person is attributed unusual properties or behavior as in Example (22). This could include strange hobbies, e.g. staring at an empty wall for hours, preferences, e.g. enjoying other people’s failures, or beliefs, e.g. believing in fairy tales. It could also be claimed that the targeted person brings about unusual situations or events, e.g. causing others run away screaming. They might be attributed to incite unusual behavior on the part of the speaker as in Example (23). Finally, the unusual qualities could be communicated through non-standard language, such as unusual imagery (24) or other creative phrasing (25).

The concept of unusual properties is intended to cover all cases where the abused target is meant to be alienated from the interlocutor(s). Therefore, this feature captures various instances of "othering," a method of stigmatizing the target as not fitting within the norms of a social group, which has frequently been observed to coincide with abusive language (Burnap and Williams, 2015).

To extract this feature with GPT models, the following prompt template was devised. The formulation was largely inspired by Wiegand et al. (2023, 16286):

> We define unusual utterances as sentences where the addressed person is attributed unusual properties or displays some unusual behavior. This could be strange hobbies, preferences or beliefs. The addressed person could also cause unusual situations or events or unusual behavior on the part of the speaker. The unusual property may also be conveyed by the usage of non-standard language, i.e., unusual imagery or some creative wording. The intention of the speaker is to alienate the addressed person from the reader. Based on this definition, does the following sentence describe an unusual property, behavior or situation? "{SENTENCE}" Answer either "yes" or "no"; no other answers are permitted.

### Few-Shot Feature Extraction: Unusual Properties

Consistent with previous methods, for the few-shot approach, the examples were selected in a manner identical to other linguistic features and combined with the prompt already devised during the zero-shot approach:

> We define unusual utterances as sentences where the addressed person is attributed unusual properties or displays some unusual behavior. This could be strange hobbies, preferences or beliefs. The addressed person could also cause unusual situations or events or unusual behavior on the part of the speaker. The unusual property may also be conveyed by the usage of non-standard language, i.e., unusual imagery or some creative wording. The intention of the speaker is to alienate the addressed person from the reader. Based on this definition, does the following sentence describe an unusual property, behavior or situation? "{SENTENCE}" Answer either "yes" or "no"; no other answers are permitted. Here are examples:
>
> Sentence: "You have never seen a feather duster in your life, have you." Answer: "yes"
>
> Sentence: "You don't like tidying up." Answer: "no"
>
> Sentence: "You will make me want to do very nasty things." Answer: "yes"
>
> Sentence: "It seems university is not the right fit for you." Answer: "no"
>
> Sentence: "You will cause anyone to scream and run away." Answer: "yes"
>
> Sentence: "Punctuality is not one of your strengths." Answer: "no"
>
> Sentence: "Your brain bid farewell to you a long time ago." Answer: "yes"
>
> Sentence: "I'm not pleased with your decision to bring all these individuals along." Answer: "no"
>
> Sentence: "You wouldn't know honesty if it passed you in the street." Answer: "yes"

Sentence: “It’s fantastic that you opened up to him.” Answer: “no”

The examples “You have never seen a feather duster in your life, have you,” “You don’t like tidying up,” “You will make me want to do very nasty things,” “You will cause anyone to scream and run away,” “Your brain bid farewell to you a long time ago,” and “You wouldn’t know honesty if it passed you in the street” were sourced from the annotation guidelines (Wiegand et al., 2023). The remaining examples were paraphrases of sentences from the dataset *Euphemistic Abuse*. The original sentences were: “I’m afraid, graduate school just isn’t working for you,” “Being on time is not your strong suit,” “I don’t appreciate that you brought all these people with you,” and “How wonderful that you shared with him.”

Results of the feature extraction are discussed in Chapter 6.

### 5.4.3. Comparisons

Wiegand et al. (2021a) present a set of linguistic features utilized for supervised classification. Some of these features were produced manually while others were generated automatically. The following specific features are produced manually:

- Figurativeness vs. literalness
- Dehumanization
- Taboos
- Evaluation vs. emotional frame of mind
- Contradiction
- Absurd images

For these features, the dataset *Comparisons* provides manually produced labels, thereby establishing a gold standard that enables evaluation.

On a random sample of 200 comparisons, Wiegand et al. (2021a) assessed the inter-annotation agreement on each manually designed feature between two annotators, one being a coauthor and the other a graduate student in computational linguistics. The resulting scores for each feature ranged from 0.63 to 1.00, which, according to Landis and Koch (1977), indicates at least a substantial level of agreement.

Wiegand et al. (2021a) also introduced linguistic features that they extracted automatically instead of manually:

- Intensity
- Frequency
- Absence of nouns and adjectives
- Similarity to explicit insults
- Emotions
- WordNet supersenses

Table 10.: Examples from the dataset *Comparisons* classified by the class label indicating whether the instance was rated as abusive or not (Abuse), the features figurativeness vs. literalness (Fig/Lit), dehumanization (Dehum), taboos (Taboo), evaluation vs. emotional frame of mind (Eval/Frame), contradiction (Contrad), and absurd images (Absurd).

| Comparison | Abuse | Fig/Lit | Dehum | Taboo | Eval/Frame | Contrad | Absurd |
|---|---|---|---|---|---|---|---|
| You walk like a giraffe. | ✓ | Fig | ✓ | ✗ | Eval | ✗ | ✗ |
| You behave like a toddler on acid. | ✓ | Fig | ✗ | ✓ | Eval | ✗ | ✓ |
| You look like you're lost. | ✗ | Lit | ✗ | ✗ | Frame | ✗ | ✗ |
| You are as modern as a caveman. | ✓ | Fig | ✗ | ✗ | Eval | ✓ | ✗ |

Table 10 displays an excerpt from the dataset. Similar to previous sections, this section examines each of the features proposed by Wiegand et al. (2021a) and describes how they were extracted with GPT models, beginning with the zero-shot approach, followed by the few-shot approach. In the few-shot approach, the method mirrored the technique applied in extracting the linguistic features for the datasets *Identity Groups* and *Euphemistic Abuse*: ten examples were presented to the GPT models, with each class represented by five examples.

### Zero-Shot Feature Extraction: Figurativeness vs. Literalness

Wiegand et al.'s (2021a, 361) dataset includes both figurative (26) and literal comparisons (27).

(26) "You sing like a dying bird." (figurative & ABUSE)

(27) "You have the face of a sad person." (literal & OTHER)

In figurative comparisons, the vehicle and the topic are fundamentally different types of entities[4]. Literal comparisons, conversely, are characterized by their reversibility. This means that the topic and vehicle of a literal comparison can switch places without significant changes in meaning. For example, "Encyclopedias are like dictionaries" can be rephrased as "Dictionaries are like encyclopedias." However, this is not applicable for "Encyclopedias are like goldmines," which would thus be classified as a figurative comparison (Wiegand et al., 2021a, 362). Additionally, literal comparisons must highlight properties that are prominent in both entities being compared. For instance, a dictionary and an encyclopedia, both types of language resources, share properties such as being organized in a specific order and containing a number of entries. These characteristics are not typically found in figurative comparisons. An encyclopedia and a goldmine only share minor properties, such as their potential profitability.

One might intuitively expect abusive comparisons to be figurative (26) and non-abusive comparisons to be literal (27). Consequently, Wiegand et al. (2021a) deemed it necessary to determine whether abusive comparisons merely coincide with figurative comparisons. Their findings indicate that literal comparisons correlate with non-abusive comparisons, but many figurative comparisons are also non-abusive. Therefore, the task of detecting abusive comparisons cannot simply be reduced to distinguishing between literal and figurative language. Wiegand et al. (2021a) had to consider additional features such as dehumanization, taboo topics, absurd images, etc.

[4] Vehicles and topics of comparisons are illustrated in Table 4.

The following prompt template was formulated to extract this feature using GPT models:

> Is the following comparison figurative? “{SENTENCE}” Answer “yes” or “no”; no other answers are permitted.

### Few-Shot Feature Extraction: Figurativeness vs. Literalness

The prompt below was developed for feature extraction using GPT models in a few-shot approach:

> Is the following comparison figurative? “{SENTENCE}” Answer “yes” or “no”; no other answers are permitted. Here are examples:
>
> Sentence: “Your singing resembles that of a bird nearing its end.” Answer: “yes”
>
> Sentence: “Your expression resembles that of someone who is sorrowful.” Answer: “no”
>
> Sentence: “Your actions mimic those of a very small child under the influence of strong substances.” Answer: “yes”
>
> Sentence: “It’s as if you possess no manners at all.” Answer: “no”
>
> Sentence: “Your words are as harsh as flames.” Answer: “yes”
>
> Sentence: “You appear to be bewildered.” Answer: “no”
>
> Sentence: “You responded as though you were a young child who’s just let go of a balloon.” Answer: “yes”
>
> Sentence: “Your arms are unexpectedly short; it’s astonishing you manage to find fitting attire.” Answer: “no”
>
> Sentence: “Talking to you is akin to dealing with an engine that can’t run smoothly.” Answer: “yes”
>
> Sentence: “Your behavior suggests you’re in dire need of some rest.” Answer: “no”

Neither Wiegand et al. (2021a) nor the respective annotation guidelines provided examples beyond those already in the dataset, so paraphrases of sentences from the dataset were used instead. The original sentences were: “You sing like a dying bird,” “You have the face of a sad person,” “You behave like a toddler on acid,” “Your manners are like well no manners really,” “Your words are like fire,” “You look like you’re lost,” “You reacted like a child who lost a balloon,” “Your arms are like so short I’m surprised you can find clothes that fit you,” “ Your conversations are like a stuttering engine,” and “You are acting like you need a nap.”

### Zero-Shot Feature Extraction: Dehumanization

A dehumanizing comparison involves directly comparing a person or their intrinsic mental or physical traits to a non-human entity. Dehumanization is commonly linked to abusive language (Mendelsohn et al., 2020). An example of a dehumanizing comparison would be (28)(Wiegand et al., 2021a, 362), which compares a physical attribute of a person, i.e., their walk, to a non-human entity, i.e., a giraffe.

(28) "You walk like a giraffe." (ABUSE)

To isolate this feature via GPT models, the subsequent prompt template was developed. The definition for *dehumanizing comparison* was derived from Wiegand et al. (2021a, 362):

> A dehumanizing comparison is defined as directly comparing a person or their inherent mental or physical attributes to a non-human entity. Based on this definition, is the following comparison dehumanizing? "{SENTENCE}" Answer either "yes" or "no"; no other answers are permitted.

**Few-Shot Feature Extraction: Dehumanization**

The subsequent prompt template was crafted to extract this feature with GPT models in the few-shot approach:

> A dehumanizing comparison is defined as directly comparing a person or their inherent mental or physical attributes to a non-human entity. Based on this definition, is the following comparison dehumanizing? "{SENTENCE}" Answer either "yes" or "no"; no other answers are permitted. Here are examples:
>
> Sentence: "Your arms resemble overcooked noodles." Answer: "yes"
>
> Sentence: "Your expression resembles that of someone who is sorrowful." Answer: "no"
>
> Sentence: "Your singing resembles that of a bird nearing its end." Answer: "yes"
>
> Sentence: "It's as if you possess no manners at all." Answer: "no"
>
> Sentence: "Your words are as harsh as flames." Answer: "yes"
>
> Sentence: "You appear to be bewildered." Answer: "no"
>
> Sentence: "Your eyes bear similarity to a bundle of potatoes." Answer: "yes"
>
> Sentence: "Your arms are unexpectedly short; it's astonishing you manage to find fitting attire." Answer: "no"
>
> Sentence: "Talking to you is akin to dealing with an engine that can't run smoothly." Answer: "yes"
>
> Sentence: "Your behavior suggests you're in dire need of some rest." Answer: "no"

Since Wiegand et al. (2021a) and the associated annotation guidelines did not include examples outside the existing dataset, paraphrased sentences from the dataset were utilized. The original sentences were: "Your arms are like cooked spaghetti," "You have the face of a sad person," "You sing like a dying bird," "Your manners are like well no manners really," "Your words are like fire," "You look like you're lost," "Your eyes are like a sack of potatoes," "Your arms are like so short I'm surprised you can find clothes that fit you," "Your conversations are like a stuttering engine," and "You are acting like you need a nap."

### Zero-Shot Feature Extraction: Taboo Topics

According to Allan and Burridge (2006, 1) a taboo is "a proscription of behavior that affects everyday life". A characteristic of abusive language is its consideration as taboo in many social contexts, employing words associated with taboo topics to express offensiveness, such as specific bodily organs, physical, and mental abnormalities, such as Examples (29) through (31) (Wiegand et al., 2021a). A significant number of terms such as "vagina" are excluded from standard abusive language detection resources (Wiegand et al., 2018), and as a result, they are not flagged as explicitly abuse because they are considered ambiguous. For instance, in medical contexts, such terms are acceptable. Allan and Burridge (2006) offer a set of semantic domains, such as sexuality, death, or diseases, which are frequently used as references for annotation guidelines (Wiegand et al., 2021a).

(29) "You eat like you have worms."

(30) "You are sweating like a dog in heat."

(31) "You make me feel like bringing up my lunch."

In order to derive this feature from GPT models, the prompt template below was crafted. The definition of "taboo" was sourced from Allan and Burridge (2006, 1):

> Taboo is a proscription of behavior that affects everyday life. Topics considered taboo include: bodies and their effluvia (sweat, snot, feces, menstrual fluid, etc.); the organs and acts of sex, micturition, and defecation; diseases, death, and killing (including hunting and fishing); naming, addressing, touching, and viewing persons and sacred beings, objects, and places; and food gathering, preparation, and consumption. Based on this definition, does the following sentence address a taboo topic? "{SENTENCE}" Answer either "yes" or "no"; no other answers are permitted.

### Few-Shot Feature Extraction: Taboo Topics

Here is the prompt template designed for the extraction of this feature employing GPT models under the few-shot method:

> Taboo is a proscription of behavior that affects everyday life. Topics considered taboo include: bodies and their effluvia (sweat, snot, feces, menstrual fluid, etc.); the organs and acts of sex, micturition, and defecation; diseases, death, and killing (including hunting and fishing); naming, addressing, touching, and viewing persons and sacred beings, objects, and places; and food gathering, preparation, and consumption. Based on this definition, does the following sentence address a taboo topic? "{SENTENCE}" Answer either "yes" or "no"; no other answers are permitted. Here are examples:
>
> Sentence: "You look as exhausted as a corpse." Answer: "yes"
>
> Sentence: "Your expression resembles that of someone who is sorrowful." Answer: "no"
>
> Sentence: "You are sweating like a marathon runner." Answer: "yes"
>
> Sentence: "It's as if you possess no manners at all." Answer: "no"

> Sentence: “Your face is as pale as a ghost.” Answer: “yes”
>
> Sentence: “You appear to be bewildered.” Answer: “no”
>
> Sentence: “You eat like a starving animal.” Answer: “yes”
>
> Sentence: “Your arms are unexpectedly short; it’s astonishing you manage to find fitting attire.” Answer: “no”
>
> Sentence: “You make me feel like throwing up.” Answer: “yes”
>
> Sentence: “Your behavior suggests you’re in dire need of some rest.” Answer: “no”

Examples beyond those already in the dataset were not provided by either Wiegand et al. (2021a) or the respective annotation guidelines, leading to the use of paraphrased sentences from the dataset. The original sentences were: “You look as tired as the dead,” “You have the face of a sad person,” “You are sweating like a dog in heat,” “Your manners are like well no manners really,” “Your face is like a ghost,” “You look like you’re lost,” “You eat like you have worms,” “Your arms are like so short I’m surprised you can find clothes that fit you,” “You make me feel like bringing up my lunch,” and “You are acting like you need a nap.”

### Zero-Shot Feature Extraction: Absurd Images

In the dataset *Comparisons*, many figurative comparisons also feature fairly absurd images such the ones seen in (32) and (33)(Wiegand et al., 2021a, 362). By absurd, Wiegand et al. (2021a) refer to vehicles that describe scenes that are very rarely or never observed in reality. Wiegand et al. (2018) noted that such images tend to be perceived as abusive.

(32) “Your input is like a baby giving their opinion on computer code.” (ABUSE)

(33) “You walk like you have three legs and four pockets full of rubble.” (ABUSE)

To extract this feature with GPT models, the following prompt template was devised. The formulation of this prompt is derived from Wiegand et al.’s understanding of absurd images (2021a, 362):

> An absurd sentence is a sentence that describes an image that is extremely rarely or never observed in real life. Based on this definition, is the following sentence absurd? “{SENTENCE}” Answer with either “yes” or “no”; no other answers are permitted.

### Few-Shot Feature Extraction: Absurd Images

This prompt template was composed for the GPT-driven extraction of this feature in the few-shot scenario:

> An absurd sentence is a sentence that describes an image that is extremely rarely or never observed in real life. Based on this definition, is the following sentence absurd? “{SENTENCE}” Answer with either “yes” or “no”; no other answers are permitted. Here are examples:
>
> Sentence: “Your contribution is akin to an infant commenting on programming.” Answer: “yes”

> Sentence: “Your expression resembles that of someone who is sorrowful.” Answer: “no”
>
> Sentence: “Your walking resembles someone with an extra limb and pockets filled with debris.” Answer: “yes”
>
> Sentence: “It’s as if you possess no manners at all.” Answer: “no”
>
> Sentence: “Your method is comparable to a bus colliding with a fuel truck.” Answer: “yes”
>
> Sentence: “You appear to be bewildered.” Answer: “no”
>
> Sentence: “You cook as though you’ve interpreted the recipe in reverse.” Answer: “yes”
>
> Sentence: “Your arms are unexpectedly short; it’s astonishing you manage to find fitting attire.” Answer: “no”
>
> Sentence: “Your politeness is comparable to an elephant rampaging through a porcelain store.” Answer: “yes”
>
> Sentence: “Your behavior suggests you’re in dire need of some rest.” Answer: “no”

Since Wiegand et al. (2021a) and the respective annotation guidelines did not offer examples beyond those present in the dataset, sentences from the dataset were paraphrased for use. The original sentences were: “Your input is like a baby giving their opinion on computer code,” “You have the face of a sad person,” “You walk like you have three legs and four pockets full of rubble,” “Your manners are like well no manners really,” “Your approach is like a bus crashing into a [sic] oil tanker,” “You look like you’re lost,” “You cook like you read the instructions backwards,” “Your arms are like so short I’m surprised you can find clothes that fit you,” “Your manners are like a bull in a china shop,” and “You are acting like you need a nap.”

### Zero-Shot Feature Extraction: Contradictions

A recurring construction in comparisons involves contradictions as in (34) and (35)(Wiegand et al., 2021a, 362). These generally occur when the characteristic of the comparison, e.g. “smart,” is contrary to the prototypical characteristics associated with the vehicle, e.g. a neanderthal is typically considered to be simpleminded rather than smart. Such contradicting comparisons, categorized as a sub-type of sarcasm, could be interpreted as abusive.

(34) “You are as thin as an elephant.” (ABUSE)

(35) “You are as smart as a neanderthal” (ABUSE)

This feature was extracted using GPT models by devising the following prompt template:

> Is there a contradiction in the following sentence? “{SENTENCE}” Answer either “yes” or “no”; no other answers are permitted.

### Few-Shot Feature Extraction: Contradictions

The ensuing prompt template was devised to extract this feature using GPT models and a few-shot approach:

> Is there a contradiction in the following sentence? "{SENTENCE}" Answer either "yes" or "no"; no other answers are permitted. Here are examples:
>
> Sentence: "You are as slender as a mammoth." Answer: "yes"
>
> Sentence: "You are as slender as a person emerging from a drastic fast." Answer: "no"
>
> Sentence: "You are as enlightened as a prehistoric human." Answer: "yes"
>
> Sentence: "You are as ravenous as a predator." Answer: "no"
>
> Sentence: "You are as contemporary as a Stone Age hunter." Answer: "yes"
>
> Sentence: "Your dialogues are as tedious as watching paint dry." Answer: "no"
>
> Sentence: "Your etiquette is as polished as an ape's." Answer: "yes"
>
> Sentence: "You appear as weary as someone pulling a double shift." Answer: "no"
>
> Sentence: "Your advancement is as if you're moving in reverse." Answer: "yes"
>
> Sentence: "You are as unhearing as a piece of wood." Answer: "no"

Due to the lack of examples beyond the dataset from Wiegand et al. (2021a) and the respective annotation guidelines, paraphrases of dataset sentences were utilized. The original sentences were: "You are as thin as an elephant," "You are as thin as someone who's just done an extreme diet," "You are as smart as a neanderthal," "You are as hungry as the wolf," "You are as modern as a caveman," "Your conversations are as dull as dishwater," "Your manners are as refined as a monkey's," "You look as tired as an overtime worker," "Your progress is like you're stepping backwards," and "You are as deaf as a doorpost."

### Zero-Shot Feature Extraction: Evaluation vs. Emotional Frame of Mind

Even though all instances from the dataset are negative in polarity, they vary in the type of sentiment expressed. On one hand, there are evaluative comparisons (36)(Wiegand et al., 2021a, 362), where the author of a comparison negatively evaluates a specific property of the target person, typically criticizing their behavior or physical appearance, e.g. being overweight (36). Such evaluative comparisons are more likely to be viewed as abusive expressions. On the other hand, there are comparisons where the author describes the emotional state of the target (37)(Wiegand et al., 2021a, 362). Given that all instances from the dataset *Comparisons* are negative, typical emotional states include pain, sorrow, exhaustion, or shock, as in (37). In these cases, the author does not necessarily evaluate the target. For example, stating that someone is suffering does not imply criticism but rather concern. Such comparisons are rarely perceived as abusive.

(36) "You look like an overfed cat." (ABUSE)

(37) "You look like a shocked cat." (OTHER)

To capture this feature using GPT models, the following prompt template was constructed. The definitions were inspired by Wiegand et al. (2021a, 362):

> An evaluative comparison involves the author negatively assessing a specific aspect of the addressed person (the addressed person is determined by the second-person pronoun "you"), often by criticizing their behavior or outward appearance. On the other hand, a non-evaluative comparison describes the emotional state of the addressed person without necessarily passing judgment. Based on these definitions, is the following sentence evaluative? "{SENTENCE}" Answer either "yes" or "no"; no other answers are permitted.

### Few-Shot Feature Extraction: Evaluation vs. Emotional Frame of Mind

Outlined below is the prompt template used for the few-shot approach to extract this feature:

> An evaluative comparison involves the author negatively assessing a specific aspect of the addressed person (the addressed person is determined by the second-person pronoun "you"), often by criticizing their behavior or outward appearance. On the other hand, a non-evaluative comparison describes the emotional state of the addressed person without necessarily passing judgment. Based on these definitions, is the following sentence evaluative? "{SENTENCE}" Answer either "yes" or "no"; no other answers are permitted. Here are examples:
>
> Sentence: "You resemble a feline that's been fed too much." Answer: "yes"
>
> Sentence: "You appear to be bewildered." Answer: "no"
>
> Sentence: "Your actions mimic those of a very small child under the influence of strong substances." Answer: "yes"
>
> Sentence: "Your expression resembles that of someone who is sorrowful." Answer: "no"
>
> Sentence: "Your contribution is akin to an infant commenting on programming." Answer: "yes"
>
> Sentence: "Your response brings to mind a startled cat." Answer: "no"
>
> Sentence: "You are as enlightened as a prehistoric human." Answer: "yes"
>
> Sentence: "It seems as though you're burdened with heavy loads on your back. " Answer: "no"
>
> Sentence: "Your etiquette is as polished as an ape's." Answer: "yes"
>
> Sentence: "Your reaction was as if you were found in a compromising situation without your pants. " Answer: "no"

Wiegand et al. (2021a) and the respective annotation guidelines did not provide examples beyond the dataset, so paraphrased sentences from the dataset were used instead. The original sentences were: "You look like an overfed cat," "You look like you're lost," "You behave like a toddler on acid," "You have the face of a sad person," "Your input is like a baby giving their opinion on computer code," "Your reaction reminds me of a shocked cat," "You are as smart as a neanderthal," "You are acting like you are carrying weights on your back," "Your manners are as refined as a monkey's," and "You reacted like you got caught with your trousers down."

### Zero-Shot Feature Extraction: Intensity

This feature represents the first to be automatically generated by Wiegand et al. (2021a). In their 2018 study, Wiegand et al. established a correlation between high polar intensity in language and its abusive nature (38)(Wiegand et al., 2021a, 363). To quantify the polar intensity of a comparison, Wiegand et al. (2021a) utilized the intensity lexicon from their 2018 research, ranking each comparison based on the average intensity score of the words it contained. The lexicon categorizes words along a spectrum from very positive to very negative, placing polar intensive words at both extremes of this spectrum.

(38) "Your eyes are like doorways into hell itself." (ABUSE)

The prompt devised to extract this feature with GPT models was as follows:

> Does the following sentence use polar intense words? "{SENTENCE}" Answer either "yes" or "no"; no other answers are permitted.

The few-shot approach was skipped for the linguistic features intensity, frequency, absence of nouns and adjectives, similarity to explicit insults, emotions, and WordNet supersenses. Only the zero-shot prompting method was used for those features that Wiegand et al. (2021a) extracted automatically and without human annotators. The decision to utilize zero-shot prompting only was based on its anticipated sufficiency for tasks akin to comparing word lists. Employing few-shot prompting with GPT models was deemed unnecessary and overly time-consuming for a task expected to yield uninspiring results. Additionally, the time required to create the extensive number of examples needed was not available. Specifically, the features "emotions" and "WordNet supersenses" would have required a significant number of examples: ten for each of the eight emotions identified by Mohammad and Turney (2013) and another ten for each of the 45 WordNet supersenses (Miller et al., 1990), totaling 530 examples just for these two features.

### Zero-Shot Feature Extraction: Frequency

High polar intensity in language may not only be conveyed by inherently polar words such as "hell," but also through comparisons to special items. Wiegand et al. (2021a) posit that such items typically share the property of being infrequent in general text corpora. Words like "bakelite" or "kazoo" are not polar expressions themselves but are rare in text corpora. When used in comparisons like (39) through (40)(Wiegand et al., 2021a, 363), these comparisons are perceived as extreme and therefore likely to be abusive. For their experiments, Wiegand et al. (2021a) determined the word frequency using the North American News Corpus (Graff, 1995) and sorted each comparison based on its least common word.

(39) "You are as modern as bakelite." (ABUSE)

(40) "You laugh like a kazoo." (ABUSE)

Employing GPT models to extract this feature, the following prompt template was set up:

> Does the following sentence use rare, infrequent words? "{SENTENCE}" Answer either "yes" or "no"; no other answers are permitted.

### Zero-Shot Feature Extraction: Absence of Nouns and Adjectives

In abusive comparisons, the imagery typically requires concrete nouns as the vehicle. Additionally, abusive comparisons may necessitate adjectives to convey high polar intensity or a negative evaluation. This leads to the conclusion that the absence of these two parts of speech is likely indicative of a non-abusive comparison such as Examples (41) through (42)(Wiegand et al., 2021a)(Wiegand et al., 2021a, 363). Wiegand et al. extracted this feature with the help of part-of-speech tagging.

(41) “You look like you’re lost.” (OTHER)

(42) “You move like you’re hurt.” (OTHER)

To extract this feature in GPT models, the following prompt template was established:

> Does the following sentence use nouns or adjectives? “{SENTENCE}” Answer either “yes” or “no”; no other answers are permitted.

### Zero-Shot Feature Extraction: Similarity to Explicit Insults

Although the comparisons in the dataset curated by Wiegand et al. (2021a) do not contain any explicitly abusive words, employing a lexicon of such words may still aid in classification. The *ISHate* dataset with explicit labels could not be utilized here because it was discovered subsequently to these experiments. Wiegand et al. (2018) acknowledge that the boundary between explicitly and implicitly abusive language is not well-defined, noting that the dataset contains ambiguous abusive words that exhibit significant semantic resemblance to abusive words from a lexicon. Wiegand et al. (2021a) utilized the lexicon from their 2018 study, computed a centroid embedding vector of its entries, and ranked the instances from the dataset based on their semantic similarity to this centroid. The comparison was depicted by the embedding vector of the word within that comparison which had the greatest similarity to the centroid. For the embeddings, Wiegand et al. (2021a) selected the FastText embeddings (Joulin et al., 2017) trained on Common Crawl.

The extraction of this feature with GPT models utilized the following prompt template:

> Does the following sentence use abusive words? “{SENTENCE}” Answer either “yes” or “no”; no other answers are permitted.

### Zero-Shot Feature Extraction: Emotions

In response to the findings by Rajamanickam et al. (2020), which highlighted a correlation between abusive language and emotions, Wiegand et al. (2021a) employed the lexicon developed by Mohammad and Turney (2013) in order to study the emotions associated with abusive and non-abusive comparisons. Mohammad and Turney’s lexicon categorizes eight emotions associated with commonly used English words: joy, sadness, anger, fear, trust, disgust, anticipation, and surprise. A single word may be linked to multiple emotion categories. In their study, Wiegand et al. (2021a) analyzed comparisons by identifying the emotion categories from the aforementioned lexicon associated with the words in each comparison. They discovered that the emotion of disgust tends to correlate with abusive comparisons, whereas trust is more often found in non-abusive instances.

The attempt to automate the extraction of this feature using GPT models faced challenges due to the initial design of the prompts, which were intended for binary classification. To adapt, a distinct prompt was crafted for each emotion category to maintain consistency in the approach. These prompts are structured to identify a specific emotion from a given sentence and to classify the presence of that emotion as either “yes” or “no”. Algorithm 3 demonstrates the approach.

**Algorithm 3** Emotion Identification Algorithm

```
emotions ← [“joy”, “sadness”, “anger”, “fear”, “trust”, “disgust”, “anticipation”, “surprise”]
SENTENCE ← Instance from the dataset
for emotion in emotions do
    prompt_template ← ‘Identify the emotions associated with the following sentence: “{SENTENCE}” Is one of the identified emotions {emotion}? Answer either “yes” or “no”; no other answers are permitted.’
    chat_completion ← Get GPT model response using:
        prompt → prompt_template.format(emotion=emotion, SENTENCE=SENTENCE)
end for
```

**Zero-Shot Feature Extraction: WordNet Supersenses**

In their research, Wiegand et al. (2021a) investigated the role of WordNet supersenses (Miller et al., 1990), which comprise 45 broad semantic categories, e.g. verbs related to fighting and nouns related to body parts. These supersenses have proven to be effective in various linguistic tasks, such as sentiment analysis (Flekova and Gurevych, 2016).

Wiegand et al. (2021a) aimed to determine if specific supersenses are associated with abusive language. They identified that comparisons involving animals, food, and body parts are often characteristic of abusive language. Conversely, comparisons that relate to weather phenomena, events, temporal expressions, and acts generally suggest non-abusive contexts. Additionally, verbal categories such as motion, perception, and states often co-occur with non-abusive comparisons. From these observations, Wiegand et al. (2021a) concluded that comparisons involving abstract concepts tend to be non-abusive, whereas those involving more tangible entities, such as animals or food, are more likely to be abusive.

When extracting this feature with GPT models, similar challenges were encountered as with the emotions feature. To address these, a distinct prompt was crafted for each of the 45 supersenses to ensure consistency in the approach. The prompts are designed to identify a specific supersense from a given sentence and to classify the presence of that supersense as either yes or no. The method is illustrated by Algorithm 4, which closely resembles Algorithm 3.

## 5.5. Combination of GPT Completions with Supervised Machine Learning

This section discusses the integration of GPT-generated completions with rule-based and logistic regression classifiers to enhance the detection of implicit hate speech. It details the development of classifiers that utilize the previously described linguistic features extracted through automated processes, outlining both the method and the specific algorithms employed.

**Algorithm 4** Supersense Analysis Algorithm

```
supersenses ← [“unique beginner for nouns”, “…”, “verbs of raining, snowing, thawing, thundering”]
SENTENCE ← Instance from the dataset
for supersense in supersenses do
    prompt_template ← ‘Does the following sentence use any {supersense}? “{SENTENCE}” Answer either “yes” or “no”; no other answers are permitted.’
    chat_completion ← Get GPT model response using:
        prompt → prompt_template.format(supersense=supersense, SENTENCE=SENTENCE)
end for
```

## 5.5.1. Rule-Based Classifier for the Dataset *Identity Groups*

Initially, a simple rule-based classifier (Wiegand et al., 2022) was implemented, utilizing outputs derived from the previously described automated feature extraction processes. The underlying logic and specific conditions applied are detailed in Algorithm 5. To qualify as hate speech, a sentence must be inherently non-episodic. Additionally, it has to either portray the targeted identity group in the role of a perpetrator or link the group with non-conformist views.

**Algorithm 5** Determine if Implicitly Abusive

```
function IsImplicitlyAbusive($aspect$, $perpetrator$, $nonConformistic$)
    $prediction$ ← “OTHER”
    if $aspect$ = “NON-EPISODIC” and ($perpetrator$ = “YES” or $nonConformistic$ = “YES”) then
        $prediction$ ← “ABUSE”
    end if
    return $prediction$
end function
```

Example (43)(Wiegand et al., 2022, Dataset) is episodic and is therefore classified as other. Example (44)(Wiegand et al., 2022, Dataset), while non-episodic, does not depict the target as either a perpetrator or a non-conformist, and is thus also classified as other. In contrast, Example (45)(Wiegand et al., 2022, Dataset) is non-episodic and portrays the target as a perpetrator, resulting in a classification of abuse. Similarly, Example (46)(Wiegand et al., 2022, Dataset), which is non-episodic and presents the target as a non-conformist, is also labeled abuse.

(43) “Jews censure Miley Cyrus for not raising Gypsy apartheid issue on ‘Gypsy Heart’ tour.” (episodic, non-perpetrator, non-conformistic, other)

(44) “All women grapple with the voice of self- doubt [sic] in one way or another.” (non-episodic, non-perpetrator, conformistic, other)

(45) “Lesbians pressure straight women into being lesbians all the time.” (non-episodic, perpetrator, conformistic, abuse)

(46) “Muslims lack basic morality bcoz quran [sic] lacks morality lessons.” (non-episodic, perpetrator, non-conformistic, abuse)

These linguistic features specifically relate to sentences in which identity groups are represented as the agent. This particular orientation makes the classifier suited for application to the dataset *Identity Groups* only. Unfortunately, the datasets *Euphemistic Abuse*, *Comparisons*, and *ISHate* were not annotated with these specific features, namely, the non-episodic aspect, perpetrator, and non-conformist, so the rule-based classifier could not be applied to these datasets.

The structure of this classifier mirrors that of the one outlined in the study by Wiegand et al. (2022), with the only difference that it employs outputs from GPT-driven feature extraction rather than relying on manually extracted features. The features used in this master's thesis were extracted through the methods previously described as zero-shot and few-shot approaches. The effectiveness of this classifier is evaluated by comparing its results to those reported in Wiegand et al. (2022) in Chapter 6.

### 5.5.2. Logistic Regression

With the multitude of linguistic features and complex automated extraction methods, it is easy to lose focus on the primary goal, namely the detection of implicit hate speech. Each identified linguistic feature brings this objective closer to realization. To investigate whether the GPT-driven feature extraction contributed to the detection of implicit hate speech, four logistic regression models for binary classification were trained. The models used the same algorithmic setups, however, they had different training scenarios:

- The first model utilized completions from our previously established baseline, described in Section 5.3, i.e., the responses to the five prompts with the markers hateful, abusive, offensive, toxic, and insulting. These completions were used as feature inputs. This training scenario was employed to determine if there is an added benefit to combining all five markers, rather than focusing on the F1 score of a single marker or using the predictions of one marker alone. The objective was to assess the predictive power of the combined markers and to compare the results to the baseline and the next three logistic regression models.

- The second model was trained using completions derived from the automated extraction of linguistic features described in Section 5.4.

- The third model was a blend of the previous two. It combined completions from both the automated linguistic feature extraction and the baseline prompts, i.e., the responses to the five prompts with the markers hateful, abusive, offensive, toxic, and insulting.

- Lastly, the fourth model used completions from the automated extraction of linguistic features and the completion from the prompt with the marker offensive. This choice was made due to its consistently good, though not always optimal, results with both GPT-3.5 and GPT-4.

These experiments were conducted once with completions from the zero-shot prompting approach and again with completions from the few-shot prompting approach.

This method was implemented with Python 3.12.0 (Van Rossum and Drake, 2009), utilizing the pandas library (The Pandas Development Team, 2024) for data manipulation and the scikit-learn library (Pedregosa et al., 2011) to preprocess data, train the models, and evaluate their performance. LogisticRegression, with L2 regularization as per the default setting, was used for logistic regression. OneHotEncoder converted categorical variables into a numerical format suitable for machine learning algorithms. StratifiedKFold provides a stratified K-Folds cross-validator that preserves the percentage of samples for each class. Precision_score, recall_score, and f1_score were used to assess the models' precision, recall, and F1 score.

The logistic regression models were trained with cross-validation of five folds. For experiments with the dataset *Identity Groups*, the scikit-learn library LogisticRegressionCV with built-in cross-validation was used. The seed of the parameter random_state was set to a fixed number to ensure reproducibility. To maintain scientific rigor, the test sets from each fold were preserved. This means that the five folds used for the logistic regression were set randomly, but the same random seed was used for each run to maintain consistency across experiments.

The presented logistic regression models for the dataset *Euphemistic Abuse* utilized the same data folds as those employed by Wiegand et al. (2023). This was done to maintain consistency and comparability with their analytical methods. Wiegand et al. (2023) organized the folds to ensure that euphemistic abusive sentences for a given cue phrase were confined to a single fold. Cue phrases, selected by the authors, were provided to crowdworkers and consisted of explicitly abusive sentences such as "You are ugly," which the crowdworkers were tasked with paraphrasing. Consequently, the test data always included euphemistic abuse derived from unseen cue phrases. In contrast, randomly assigning sentences to folds would be less stringent, as euphemistic abuse could then originate from cue phrases encountered during training.

Similarly, in the presented experiments conducted with the *Comparisons* dataset, the same data folds used by Wiegand et al. (2021a) were utilized. This ensured consistency and comparability with Wiegand et al.'s (2021a) analytical methods. They organized the folds to include mutually exclusive patterns[5], so the test instances always contained patterns not observed in the training data.

## 5.6. Evaluation

Before presenting the results, the evaluation metrics and methods need to be explained. Prior to initiating the linguistically informed approach outlined in Section 5.4, a more direct route was explored, i.e., a baseline for direct implicit hate speech detection with GPT models was established. As discussed in Section 5.3, the objective was to assess how well the GPT models could identify implicit hate speech directly. The results obtained from this initial test served as a baseline against which the results of subsequent experiments would be compared, establishing the benchmark to surpass.

This baseline was evaluated as follows. One of the linguistic markers for hate speech, hateful, abusive, offensive, toxic, or insulting, was designated as the positive class, and other as the negative class for the purpose of binary classification. Precision and recall were computed for both the positive and negative classes. These calculated values were then averaged to derive a unified precision and recall metric, which was used to calculate the average macro F1 score. This evaluation method was adopted to ensure consistency and comparability with the analytical procedures employed by Wiegand et al. (2021a, 2022, 2023). By mirroring their method of F1 score calculation, a direct and reliable basis for comparing the outcomes of the research with their findings was established.

[5] Patterns of comparisons are illustrated in Table 4.

Furthermore, this metric treats all classes equally, regardless of their size, and is more appropriate when performance on all classes is equally important. In contrast, other types like micro-averaging pool decisions for all classes into a single confusion matrix, and precision and recall are computed from this combined data. This approach is dominated by the more frequent classes due to the aggregation of all counts (Jurafsky and Martin, 2024c).

When the results were unexpected, as will be described in Sections 6.1 and 6.2, a manual inspection was performed; recall, precision, and wrongly predicted examples were investigated.

The GPT-driven feature extraction was evaluated in the same way. The presence of the relevant feature was designated as the positive class, and its absence as the negative class for the purpose of binary classification. The macro average F1 score was then calculated as previously described. By deriving the macro average F1, the effectiveness of the automated extraction of individual features can be evaluated.

However, this analysis alone was not deemed sufficient. More critically, it is necessary to determine whether the features genuinely aid in detecting implicit hate speech and if the use of GPT models enhances this capability. To explore these questions, logistic regression was utilized, as Jurafsky and Martin (2024b) have highlighted its effectiveness in identifying relationships between specific features and outcomes.

In order to determine whether the GPT-driven feature extraction was helpful for the task of implicit hate speech detection, the results of logistic regression models trained on GPT-generated data were compared to the results of models trained on manually extracted data, i.e., the models presented by Wiegand et al. (2021a, 2022, 2023). This is also why the aforementioned macro average F1 score was selected for the evaluation. By using this specific metric, the results of the experiments could be compared to Wiegand et al.'s (2021a; 2022; 2023) results.

# 6. Results

This chapter presents the results of the previously described experiments. It begins by discussing the attempt to replicate a task proposed by Ocampo et al. (2023) using the *ISHate* dataset and GPT models, which involves classifying text into three categories: non-hate speech, implicit hate speech, and explicit hate speech. Next, the baseline for detecting implicit hate speech with GPT models is discussed. Following this, the chapter explores the outcomes of the automated extraction of linguistic features of implicit hate speech using GPT models. It then examines the effectiveness of rule-based classifiers with data generated by GPT models and concludes with an analysis of the results from logistic regression models trained with features generated by GPT models.

## 6.1. *ISHate*: Classification Into Non-Hate Speech, Explicit and Implicit Hate Speech with GPT Models

As discussed in Section 5.3, an attempt was made to replicate one of the tasks proposed by Ocampo et al. (2023) on the *ISHate* dataset, namely the classification between non-hate speech, implicit, and explicit hate speech, using GPT models. Table 11 reports on the results. DeBERTa refers to the model developed by Ocampo et al. (2023), which is a DeBERTa V3-based classifier, fine-tuned for hate speech detection on the White Supremacist Forum dataset (de Gibert et al., 2018). This model utilizes a one-layer feed-forward network and was optimized on the *ISHate* dataset. Although Ocampo et al. (2023) tested other models, the DeBERTa V3 achieved the best results.

As seen in Table 11, few-shot prompts for GPT-3.5 were more successful than zero-shot prompts on non-hate speech and implicit hate speech detection. However, GPT-3.5 was generally outperformed by GPT-4. While GPT-4's performance in non-hate speech detection slightly improved from zero-shot to few-shot prompting, its performance decreased for both explicit and implicit hate speech classes. Notably, DeBERTa significantly outperformed all GPT variants across all three classes. In particular, for recognizing implicit forms of hate speech, DeBERTa surpassed the best-performing GPT variant, GPT-4 in the zero-shot approach, by 23.24%. Nevertheless, DeBERTa itself did not perform exceptionally well in this task, achieving an F1 score of only 44.20%.

The performance of the GPT models on the *ISHate* dataset is notably poor compared to other datasets. This issue appears to stem from two main concerns with *ISHate*. Firstly, the dataset showed inconsistencies in labeling, as seen in Examples (1) to (3) (Ocampo et al., 2023). These examples could be considered explicit hate speech, yet they were labeled as non-hate speech.

(1) "little black kid gets his ass kicked - YouTube Because I always enjoy these videos ."
Label given in *ISHate*: Non-Hate Speech

(2) “88 In the end is victory” Label given in *ISHate*: Non-Hate Speech

(3) “Whiteteacher puts Negro in his place [sic] !” Label given in *ISHate*: Non-Hate Speech

Conversely, Examples (4) to (6) (Ocampo et al., 2023), which could be regarded as implicit hate speech, were labeled as explicit hate speech.

(4) “Immigrants take our jobs, they take our country.” Label given in *ISHate*: Explicit Hate Speech

(5) “Why would you want to be respectful of women if all women are the same?” Label given in *ISHate*: Explicit Hate Speech

(6) “Feminism is a violent ideology that is hostile to men.” Label given in *ISHate*: Explicit Hate Speech

The annotations within the dataset seem questionable, leading to a belief that the gold standard was compromised. Hence, responses generated by the GPT models might actually align with a more accurate classification understanding.

Secondly, Wiegand et al. (2021b) noted that datasets with a significant amount of implicitly abusive content tend to exhibit bias stemming from how the data is collected. When large datasets like *ISHate* combine data from varied domains, they become notably diverse in style, making them less coherent. Different sources contribute varying amounts of abusive content. For instance, only a small portion of the posts from one source might be abusive, while a majority from another source, like a white supremacy forum, would be considered abusive due to the predominant discussion topics being inherently racist. This variation in content and style across sources could inadvertently train classifiers to detect the style rather than the substance of abuse, affecting the overall reliability of the dataset.

Table 11.: Dataset *ISHate*: Implicit Hate Speech Classification with GPT Models

| **Approach** | **Model** | **Macro-Average F1 Non-HS** | **Macro-Average F1 Explicit HS** | **Macro-Average F1 Implicit HS** |
|---|---|---:|---:|---:|
| Zero-Shot | GPT-3.5 | 75.43 | 74.01 | 13.54 |
| Few-Shot | GPT-3.5 | 82.28 | 67.42 | 15.52 |
| Zero-Shot | GPT-4 | 85.80 | 64.08 | 20.96 |
| Few-Shot | GPT-4 | 85.49 | 52.22 | 19.01 |
| Fine-Tuned | DeBERTa | **91.30** | **85.10** | **44.20** |

To conclude, the experiments attempted to replicate the classification task on the *ISHate* dataset, categorizing speech into non-hate, explicit hate, and implicit hate using GPT models. Results showed that Ocampo et al.’s (2023) DeBERTa model significantly outperformed GPT-4, particularly in identifying implicit hate speech. The performance issues with GPT models could be linked to labeling inconsistencies and dataset bias. Some examples labeled as non-hate speech were arguably explicit hate speech, while some implicit hate speech examples were labeled as explicit. These discrepancies suggest that the *ISHate* dataset’s gold standard might be compromised, impacting the evaluation of the models’ classification performance. Additionally, dataset bias due to diverse sources and varying amounts of abusive content could have affected the reliability of the dataset and the classifiers’ performance.

## 6.2. *Identity Groups*, *Euphemistic Abuse*, and *Comparisons*: Hate Speech Detection with GPT Models

This section discusses the results from the experiments described in Section 5.3, which had the primary goal of establishing a baseline. The experiments aimed to determine how effectively GPT models can detect implicit hate speech using zero-shot and few-shot techniques, setting a performance benchmark for further experiments.

Table 12.: Datasets *Identity Groups, Euphemistic Abuse* and *Comparisons*: Implicit Hate Speech Detection with GPT Models in the Zero-Shot and the Few-Shot Prompting Approach with the Marker offensive

| Dataset | Classifier | Macro-Average F1 Zero-Shot | Macro-Average F1 Few-Shot |
|---|---|---|---|
| Identity Groups | LLaMA-2 | 52.75 | |
| Euphemistic Abuse | LLaMA-2 | 34.91 | |
| Comparisons | LLaMA-2 | 40.22 | |
| Identity Groups | Majority | 36.00 | |
| Euphemistic Abuse | Majority | 39.20 | |
| Comparisons | Majority | 33.30 | |
| Identity Groups | GPT-3.5 | 64.91 | 55.70 |
| Euphemistic Abuse | GPT-3.5 | 59.76 | 55.37 |
| Comparisons | GPT-3.5 | 58.90 | 64.95 |
| Identity Groups | GPT-4 | 81.26 | **82.59** |
| Euphemistic Abuse | GPT-4 | 76.16 | 76.38 |
| Comparisons | GPT-4 | 75.93 | 70.63 |

Table 12 shows the baseline results for direct implicit hate speech detection with LLaMA-2, GPT-3.5, and GPT-4 across various datasets. Due to limited space, only one of the five markers, hateful, abusive, offensive, toxic, or insulting, was selected for analysis. The results for the offensive marker were chosen for inclusion in Table 12 because of its consistent performance across both models and prompting approaches, providing a reliable metric for comparison. The results for all markers, models, and datasets are available in Table 23 in Appendix A.5.

As noted previously in Section 5.4 experiences with LLaMA-2 were less than satisfactory. The model was slow, prone to crashes, and often failed to respond appropriately, suggesting it might not have fully grasped the task. This observation is supported by the results presented in Table 12. To provide a clearer perspective on LLaMA-2's performance, Table 12 includes results from a majority classifier, as provided by Wiegand et al.(2021a, 364, 2022, 5607, 2023, 16287). This classifier predicts the most frequent class in the data for all observations. Given that using LLaMA-2 in a few-shot approach might not yield significantly better results, and with limited time available, the decision was made to discontinue using LLaMA-2 and to focus efforts on more promising experiments.

Next, the analysis shifts to GPT-3.5's ability to detect implicit hate speech. The findings indicate that GPT-3.5 generally struggled with recognizing implicitly abusive language. Surprisingly, the few-shot prompting did not lead to improvements and, for some datasets, the F1 score actually decreased. Only in the *Comparisons* dataset did the few-shot approach show improvement.

Table 13.: Dataset *Comparisons*: Implicit Hate Speech Detection with GPT-4 using Zero- and Few-Shot Prompting

| **Marker** | **F1 Zero-Shot** | **F1 10-Shot** | **F1 15-Shot** |
|---|---:|---:|---:|
| Hateful | 69.09 | 71.07 | 73.21 |
| Insulting | 71.74 | 69.30 | 71.43 |
| Offensive | 75.93 | 70.63 | 71.29 |
| Toxic | 74.76 | 71.24 | 74.37 |
| Abusive | 71.55 | 72.31 | 74.55 |

In contrast to GPT-3.5, GPT-4 demonstrated a strong ability to recognize implicit hate speech in the zero-shot approach, and showed further improvement with few-shot learning. The exception was with the *Comparisons* dataset where the F1 score decreased. A closer inspection of recall and precision revealed many false negatives in the class other, i.e., the negative class. This issue is highlighted by the low recall for the negative class, as detailed in Table 18 in Appendix A.1. A subsequent attempt with few-shot learning, providing 15 examples to the GPT model, five for the positive class and ten for the negative, resulted in improvements, as shown in Table 13. The corresponding prompt template is also available in Appendix A.1.

In conclusion, the evaluation of implicit hate speech detection reveals varying results among the models tested. LLaMA-2's performance was unsatisfactory due to technical issues and poor task comprehension, leading to its discontinuation in further experiments. GPT-3.5 generally struggled with recognizing implicitly abusive language, with few-shot prompting often not improving its performance and sometimes even reducing the F1 score. GPT-4, however, showed strong capabilities in zero-shot detection, further enhanced by few-shot learning.

## 6.3. GPT-Driven Extraction of Linguistic Features of Implicit Hate Speech

Section 5.4 discusses the development and testing of methods for automated feature extraction from text data, where the goal is to identify linguistic patterns associated with the following specific forms of implicit hate speech: implicit hate speech targeting identity groups, abusive comparisons, and euphemistic abuse. The extracted features are intended to be used in machine learning algorithms to improve the classification and understanding of implicit hate speech.

### 6.3.1. Dataset *Identity Groups*: Extraction of Linguistic Feature with GPT Models

To begin, the results of the automated feature extraction for detecting implicit hate speech targeting identity groups is presented. Section 5.4.1 discusses these features, including linguistic aspects systematically extracted using GPT models.

First, the depiction of the target as non-conformist is analyzed, focusing on whether the target group is portrayed as deviating from societal norms, which can also constitute a form of implicit hate speech. Given the complexity of this concept, two distinct approaches for the prompts were devised. In the first approach, i.e., Prompt 1 the feature was clearly defined similarly to Wiegand et al. (2022) and then a binary classification was requested. In the second approach, shown in Prompt 2 the concept of *non-conformist* was abstracted to simplify the task.

Prompt 1 Non-conformist views are sentences in which the sentiment of the person performing the action (agent) towards the person or the thing receiving the action (patient) disagrees with the sentiment of the patient. In this context, consider this sentence "{SENTENCE}" Does the author of this sentence think that {TARGET} are non-conformist? Answer either "YES" or "NO". No other answer permitted.

Prompt 2 Consider this sentence "{SENTENCE}". Does the author of this sentence have a negative sentiment towards {TARGET}? Answer either YES or NO. No other answer permitted.

Second, the non-episodic aspect identifies statements that are habitual or generalize behaviors rather than describing specific incidents. This feature aids in understanding the generalizing nature of hate speech. The corresponding prompt was:

Prompt 3 Classify the following sentence as either "EPISODIC" or "NON-EPISODIC": "{SENTENCE}". No other answer permitted. Answer "EPISODIC" if the sentence expresses information about a singular, distinctive or particular event. Answer "NON-EPISODIC" if the sentence expresses information about a reoccurring event, about something resorted to on a regular basis, something of a commonplace character, typical, normal, or standard.

Third, the portrayal of the target as a perpetrator is examined to determine whether the utterance depicts the target group as a wrongdoer, thus contributing to negative stereotyping. The prompt for this feature was:

Prompt 4 A perpetrator is a person who commits an illegal, criminal, harmful or evil act. Consider the following sentence: "{SENTENCE}". Does the author of this sentence think that {TARGET} are perpetrators? Answer either YES or NO. No other answer permitted.

Table 20 of Appendix A.2 lists all the prompts tested during the experiments described in Section 5.4.1. The above prompts were the most successful and were therefore reused in the few-shot approach. In Table 14, the best results of feature extraction using zero-shot prompting are presented. The results achieved by all prompts that were tested can be found in Table 19 in Appendix A.2.

The data demonstrates a significant advantage in favor of GPT-4, particularly in its effective handling of the perpetrator feature. Nevertheless, it is noteworthy that both GPT-3.5 and GPT-4 encountered significant challenges with the non-conformist feature, a concept that is admittedly difficult to grasp. Interestingly, when compared to the first approach described in Prompt 1, both models showed significantly improved results with the approach that abstracted the concept of non-conformist, as described in Prompt 2. Next, the feature extraction using few-shot prompting is examined, where the most effective prompts from the zero-shot approach were paired with ten examples, five examples representing each class.

Table 14.: Dataset *Identity Groups*: Best Results for the Zero-Shot Automated Extraction of Linguistic Feature of Implicit Hate Speech with GPT Models

| **Feature** | **Prompt** | **Model** | **Macro-Average F1** |
|---|---|---|---:|
| Non-Conformist | Prompt 1 | GPT-3.5 | 51.06 |
| Non-Conformist | Prompt 2 | GPT-3.5 | 58.97 |
| Aspect | Prompt 3 | GPT-3.5 | 73.95 |
| Perpetrator | Prompt 4 | GPT-3.5 | 69.31 |
| Non-Conformist | Prompt 1 | GPT-4 | 53.08 |
| Non-Conformist | Prompt 2 | GPT-4 | 62.58 |
| Aspect | Prompt 3 | GPT-4 | 83.06 |
| Perpetrator | Prompt 4 | GPT-4 | 82.24 |

Table 15 presents a comparison between the zero-shot and few-shot prompting approaches. It shows that GPT-3.5 did not benefit from the inclusion of examples in the prompts; in fact, its F1 score decreased compared to the zero-shot approach. This aligns with the findings discussed in Section 6.2, which established a baseline for implicit hate speech detection using GPT models. Conversely, GPT-4 showed a significant improvement with few-shot prompting, as its F1 score increased markedly across all three linguistic features analyzed.

Table 15.: Dataset *Identity Groups*: Feature Extraction with GPT-3.5 and GPT-4 in the Zero-Shot and the Few-Shot Approach

| **Prompting** | **Feature** | **Prompt** | **Model** | **Macro-Average F1** |
|---|---|---|---|---:|
| Zero-Shot | Non-Conformist | Prompt 2 | GPT-3.5 | 58.97 |
| Few-Shot | Non-Conformist | Prompt 2 | GPT-3.5 | 54.67 |
| Zero-Shot | Non-Conformist | Prompt 2 | GPT-4 | 62.58 |
| Few-Shot | Non-Conformist | Prompt 2 | GPT-4 | 66.28 |
| Zero-Shot | Aspect | Prompt 3 | GPT-3.5 | 73.95 |
| Few-Shot | Aspect | Prompt 3 | GPT-3.5 | 56.44 |
| Zero-Shot | Aspect | Prompt 3 | GPT-4 | 83.49 |
| Few-Shot | Aspect | Prompt 3 | GPT-4 | 84.87 |
| Zero-Shot | Perpetrator | Prompt 4 | GPT-3.5 | 69.31 |
| Few-Shot | Perpetrator | Prompt 4 | GPT-3.5 | 63.83 |
| Zero-Shot | Perpetrator | Prompt 4 | GPT-4 | 82.24 |
| Few-Shot | Perpetrator | Prompt 4 | GPT-4 | 85.06 |

### 6.3.2. Dataset *Euphemistic Abuse*: Extraction of Linguistic Feature with GPT Models

In Section 5.4.2, six linguistic features of euphemistic abuse are identified and extracted using GPT models. The first feature is negated antonyms of abusive words. This involves using phrases that are the negated forms of an abusive word, such as saying "not smart" instead of "stupid." The second feature is extremes, which includes language that employs extremes or superlatives, exaggerating qualities to the point of insult without explicit abuse. The third feature is lexicalization, referring to phrases that have become fixed expressions or idioms that might carry negative connotations. The fourth feature is opposing sentiments, identifying sentences where sentiments oppose each other, reflecting a contradictory tone often used to disguise abuse. The fifth feature is taboo topics, recognizing when discourse veers into taboo subjects, which can indicate an attempt to euphemistically express abuse while discussing normally sensitive topics. The final feature is unusual properties, which identifies when unusual or unexpected attributes are ascribed to people or groups, subtly conveying negativity or abuse. Section 5.4.2 describes each of these features in further detail and presents the prompts used to automatically extract them with GPT models.

Table 16.: Dataset *Euphemistic Abuse*: Feature Extraction with GPT-3.5 and GPT-4 in the Zero-Shot and the Few-Shot Approach

| **Prompting** | **Model** | **Feature** | **Macr-Average F1** |
|---|---|---|---:|
| Zero-Shot | GPT-3.5 | Antonym | 54.95 |
| Few-Shot | GPT-3.5 | Antonym | 49.36 |
| Zero-Shot | GPT-4 | Antonym | 65.94 |
| Few-Shot | GPT-4 | Antonym | 69.25 |
| Zero-Shot | GPT-3.5 | Extreme | 59.80 |
| Few-Shot | GPT-3.5 | Extreme | 54.79 |
| Zero-Shot | GPT-4 | Extreme | 70.72 |
| Few-Shot | GPT-4 | Extreme | 72.03 |
| Zero-Shot | GPT-3.5 | Lexicalization | 53.27 |
| Few-Shot | GPT-3.5 | Lexicalization | 55.02 |
| Zero-Shot | GPT-4 | Lexicalization | 73.59 |
| Few-Shot | GPT-4 | Lexicalization | 72.62 |
| Zero-Shot | GPT-3.5 | Opposing Sentiments | 56.79 |
| Few-Shot | GPT-3.5 | Opposing Sentiments | 55.60 |
| Zero-Shot | GPT-4 | Opposing Sentiments | 69.04 |
| Few-Shot | GPT-4 | Opposing Sentiments | 69.18 |
| Zero-Shot | GPT-3.5 | Taboo | 58.65 |
| Few-Shot | GPT-3.5 | Taboo | 58.70 |
| Zero-Shot | GPT-4 | Taboo | 71.17 |
| Few-Shot | GPT-4 | Taboo | 73.49 |
| Zero-Shot | GPT-3.5 | Unusual | 60.34 |
| Few-Shot | GPT-3.5 | Unusual | 62.00 |
| Zero-Shot | GPT-4 | Unusual | 74.27 |
| Few-Shot | GPT-4 | Unusual | 74.45 |

First GPT-3.5's capabilities at recognizing linguistic feature of euphemistic abuse is examined. As seen in Table 16, the F1 scores achieved by GPT-3.5 are not exceptionally high, generally hovering around 50 to 60%. In most cases, the zero-shot approach has slightly higher F1 scores compared to the few-shot approach. Similarly to the previous experiments, the generally lower performance in the few-shot setting across most features suggests that providing few examples does not significantly help GPT-3.5.

Next, the performance of GPT-4 on the same tasks is examined. The results demonstrate an overall improvement in F1 scores across all features compared to GPT-3.5, suggesting that GPT-4 is better equipped to handle the nuances of euphemistic language. While few-shot learning generally provides a slight advantage for certain features, indicating that additional examples help refine the model's predictions, the performance is still moderate.

### 6.3.3. Dataset *Comparisons*: Extraction of Linguistic Feature with GPT Models

In Section 5.4.3, six linguistic features of abusive comparisons are identified and extracted using GPT models. The first feature concerns absurd images, which involves comparisons that invoke bizarre or extremely unlikely scenarios, often adding a layer of ridicule or surreal critique. The second feature captures instances of contradiction, where the comparison contains elements that are logically or typically contradictory, often used for ironic or sarcastic effect. The third feature involves dehumanization, depicting a person or their traits in terms of non-human entities, typically indicating a form of devaluation or abuse. The fourth feature differentiates between evaluation and emotional frame of mind, distinguishing statements that negatively evaluate a person's attributes or actions from those that describe a person's emotional state without judgment. The fifth feature distinguishes between figurativeness and literalness, identifying comparisons where the descriptor and the topic are fundamentally different types of entities, which is figurative, versus those where they can be interchangeably used without altering the meaning, which is literal. The final feature identifies language that engages with taboos, socially or culturally prohibited topics that are often seen as offensive or inappropriate in many contexts. Section 5.4.3 provides a more detailed description of these features and explains which prompts were tested to automatically extract them using GPT models.

Table 17 presents the results of the automated extraction of these linguistic features of abusive comparisons using GPT-3.5 and GPT-4, in both the zero-shot and few-shot approaches. The findings align with those in Table 16. Using GPT-3.5, the F1 scores mostly range from 47 to 60%, indicating moderate performance. In most instances with GPT-3.5, the zero-shot approach achieves slightly higher F1 scores than the few-shot approach. This is consistent with earlier experiments, where the few-shot setting generally underperformed across most features, suggesting that providing a small number of examples does not significantly enhance GPT-3.5's effectiveness.

Next, the performance of GPT-4 on these tasks is assessed. The results indicate that few-shot prompting generally yields higher F1 scores with GPT-4. This is particularly notable in the contradiction and dehumanization features, where few-shot learning significantly improves model performance. These findings suggest that incorporating a small set of example inputs into the prompts can effectively enhance GPT-4's ability to discern complex linguistic patterns, thereby boosting its overall performance in automated feature extraction tasks.

Table 17.: Dataset *Comparisons*: Feature Extraction with GPT-3.5 and GPT-4 in the Zero-Shot and the Few-Shot Approach

| **Prompting** | **Model** | **Feature** | **Macr-Average F1** |
|---|---|---|---:|
| Zero-Shot | GPT-3.5 | Absurd | 56.88 |
| Few-Shot | GPT-3.5 | Absurd | 52.41 |
| Zero-Shot | GPT-4 | Absurd | 61.88 |
| Few-Shot | GPT-4 | Absurd | 64.10 |
| Zero-Shot | GPT-3.5 | Contradiction | 53.78 |
| Few-Shot | GPT-3.5 | Contradiction | 56.53 |
| Zero-Shot | GPT-4 | Contradiction | 68.20 |
| Few-Shot | GPT-4 | Contradiction | 83.36 |
| Zero-Shot | GPT-3.5 | Dehumanization | 52.15 |
| Few-Shot | GPT-3.5 | Dehumanization | 58.09 |
| Zero-Shot | GPT-4 | Dehumanization | 71.00 |
| Few-Shot | GPT-4 | Dehumanization | 80.09 |
| Zero-Shot | GPT-3.5 | Evaluation/Frame | 51.65 |
| Few-Shot | GPT-3.5 | Evaluation/Frame | 53.94 |
| Zero-Shot | GPT-4 | Evaluation/Frame | 56.02 |
| Few-Shot | GPT-4 | Evaluation/Frame | 55.18 |
| Zero-Shot | GPT-3.5 | Figurative/Literal | 56.14 |
| Few-Shot | GPT-3.5 | Figurative/Literal | 47.77 |
| Zero-Shot | GPT-4 | Figurative/Literal | 80.12 |
| Few-Shot | GPT-4 | Figurative/Literal | 80.80 |
| Zero-Shot | GPT-3.5 | Taboo | 60.48 |
| Few-Shot | GPT-3.5 | Taboo | 56.51 |
| Zero-Shot | GPT-4 | Taboo | 73.75 |
| Few-Shot | GPT-4 | Taboo | 74.24 |

## 6.4. Rule-Based Classifier for Implicit Hate Speech Detection Using Data Generated by GPT Models

Figure 14 shows the results of the rule-based classifiers compared to the baselines established in Section 6.2. As a reminder, the rule-based classifier discussed in Section 5.5 assesses sentences against specific linguistic criteria to determine if they constitute implicit hate speech. It examines whether a sentence is non-episodic, indicating habitual or routine descriptions rather than isolated events. Further, it checks if the sentence portrays the targeted group either as perpetrators involved in reprehensible actions or as non-conformists deviating from societal norms. If a sentence meets the non-episodic criterion and one of the latter conditions, it is classified as implicit hate speech; otherwise, it is labeled as other. Algorithm 5 shows the logic behind this simple rule-based classifier.

The first five classifiers depicted in Figure 14 are rule-based classifiers, each varying based on the data they utilize. The classifier labeled as Human Annotator is based on the basic model developed by Wiegand et al. (2022, 5606), which relies on manually extracted features. The GPT-3.5 Zero-Shot classifier uses features predicted by GPT-3.5 through zero-shot prompting, replacing annotations typically done by crowdworkers. Similarly, the GPT-3.5 Few-Shot classifier employs predictions from GPT-3.5 with the few-shot

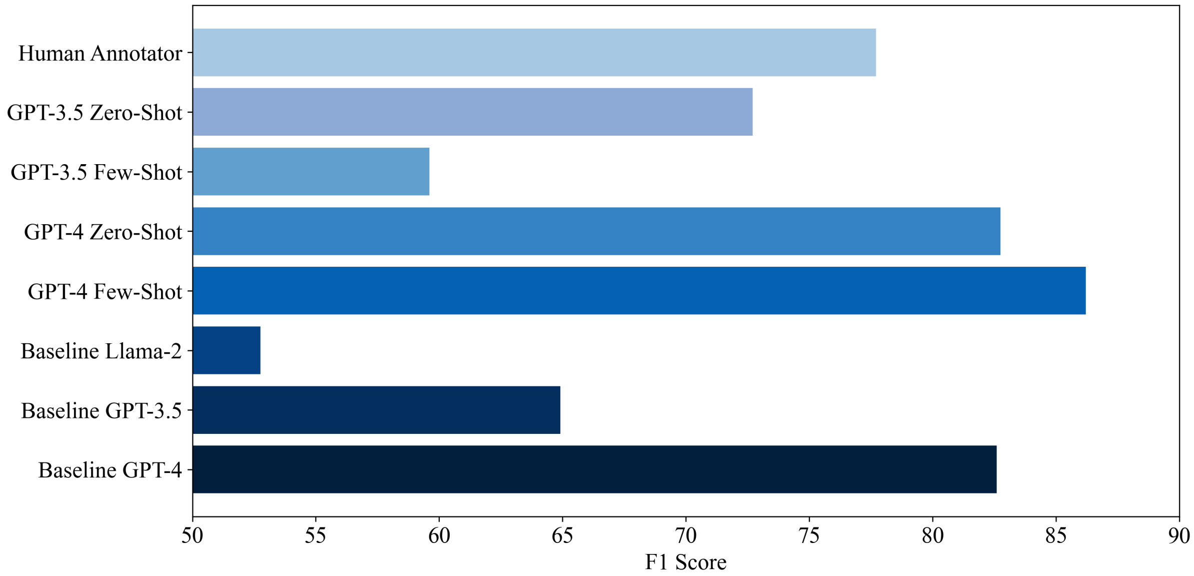


Figure 14.: Dataset *Identity Groups*: Implicit Hate Speech Detection with Rule-Based Classifiers

approach. The GPT-4 Zero-Shot classifier applies predictions from GPT-4 in zero-shot prompting, while the GPT-4 Few-Shot classifier uses predictions from GPT-4 with the few-shot approach.

These rule-based classifiers are contrasted with the baseline classifiers listed for comparison. The Baseline LLaMA-2, Baseline GPT-3.5, and Baseline GPT-4 represent the foundational standards as outlined in Section 6.2. The best results for the marker offensive were selected due to its consistent performance across both models and prompting approaches, providing a reliable metric for comparison. As illustrated in Table 23 of Appendix A.5, the offensive marker serves as a middle ground in performance, neither scoring too high nor too low, making it a good representative of the models' overall ability to detect implicit hate speech.

Figure 14 illustrates the progress made with the rule-based classifier, particularly in comparison to the classifier developed by Wiegand et al. (2022, 5606), which relied on manually extracted features and attained an average F1 score of 77.70%. Notably, the implementation of data extracted by GPT-4 through the zero-shot approach, and even more so in the few-shot approach, has enhanced the performance of this classifier by providing higher quality training data. However, the data extracted by GPT-3.5 did not yield successful outcomes, neither in the zero-shot nor in the few-shot approach. This comparison underscores the critical role of superior training data in boosting model performance.

Furthermore, contrasting the results of the rule-based classifiers with those of the baselines, shows that all classifiers outperform LLaMA-2's capabilities at direct implicit hate speech detection. Also the direct detection capabilities of GPT-4 without any features outperformed Wiegand et al.'s (2022, 5606) rule-based classifier, which relied on manually extracted features. It further shows that except for the GPT-3.5 Few-Shot model, all rule-based classifiers exceed the performance of GPT-3.5 in detecting implicit hate speech directly. Yet, only the rule-based classifiers utilizing data produced by GPT-4 significantly surpass not only the direct detection capabilities of GPT-4 but also the classifier using manually extracted features. This demonstrates that the three linguistic features, aspect, non-conformist, and perpetrator roles, are predictive of implicit hate speech. Additionally, this data suggests that substituting human annotators with GPT-4 not only is feasible but also leads to superior results. Details of the F1 scores for each classifier are available in Table 21 in Appendix A.3.

## 6.5. Logistic Regression Trained on Data Generated by GPT Models

The goal of the experiments described in Section 5.5 was to determine if the features extracted automatically by GPT-3.5 and GPT-4 were predictive of implicitly abusive language. To this end, four different logistic regression models were trained, each with the same experimental setup but using different training data:

- The first model was trained using completions derived from the automated extraction of linguistic features described in Section 6.3. In the upcoming figures, this model is called Ling. Feat.

- The second model used completions from the previously established baseline, described in Section 6.2, namely the responses to the five prompts marked hateful, abusive, offensive, toxic, and insulting. These completions served as feature inputs and the model is labeled Completions DHSD (Direct Hate Speech Detection) in the figures.

- The third model combined completions from both the automated linguistic feature extraction and the baseline prompts, referred to as Ling. Feat. + Completions DHSD in the figures.

- The fourth model utilized completions from the automated extraction of linguistic features and the response from the prompt marked offensive. This choice was based on its consistently good, though not always optimal, performance with both GPT-3.5 and GPT-4. This model is named Ling. Feat. + Completion “offensive” in the figures.

The figures also include the baseline established in Section 6.2, displaying values corresponding to the F1 score for the offensive marker in either zero-shot or few-shot prompting approach with GPT-3.5 or GPT-4, depending on the experiment being analyzed. As previously explained the marker offensive was selected due to its consistent performance across both models and prompting approaches. Details of the F1 scores for each model across all figures are available in Table 22 in Appendix A.4.

In addition, the figures feature a dotted horizontal line labeled LRT-MGF, which stands for Logistic Regression Trained on Manually Generated Features. This line represents the F1 scores reported by Wiegand et al. (2021a, 364, 2023, 16287). Unlike the experiments presented here, which use data from GPT models, their studies utilized data collected by human annotators. Additionally, for comparison, a logistic regression model was trained using the gold standard data from the *Identity Groups* dataset, which was produced by crowdworkers. The horizontal dotted line in Figures 15 and 16 represents the F1 score from this logistic regression model.

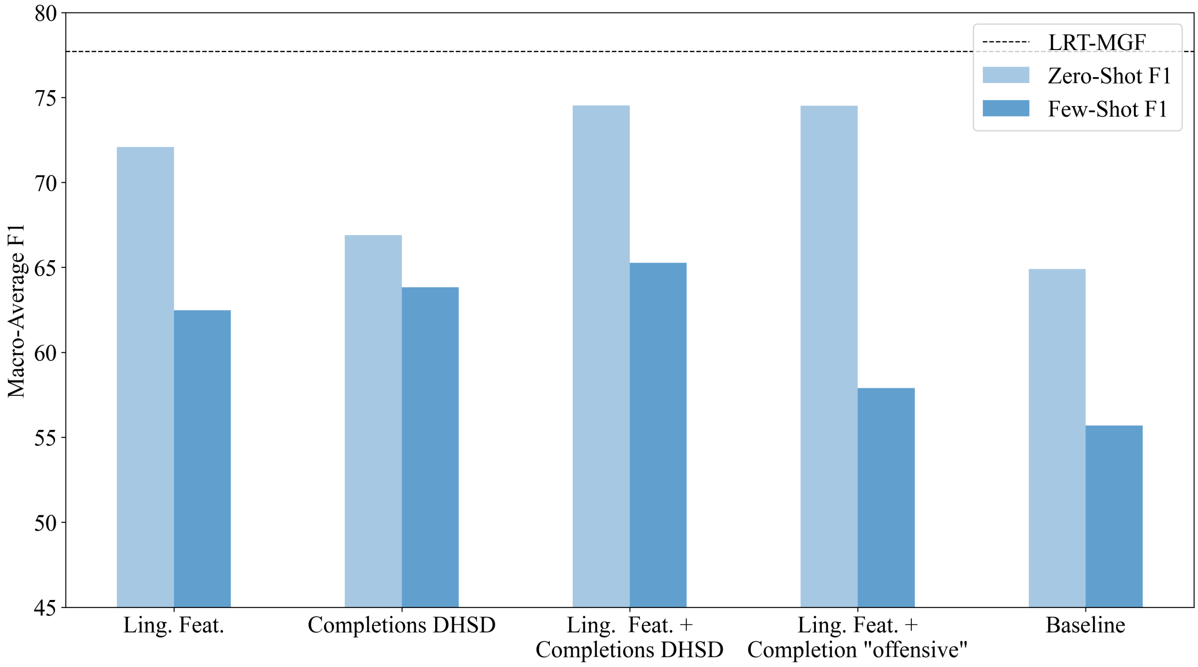


Figure 15.: Dataset *Identity Groups*: Logistic Regression Classifiers Trained on Data Generated by GPT-3.5

### 6.5.1. Dataset *Identity Groups*: Logistic Regression Trained on Data Generated by GPT Models

Figure 15 displays the results from logistic regression models trained with GPT-3.5 data on the *Identity Groups* dataset using both zero-shot and few-shot prompting. The few-shot approach shows a reduction in F1 scores compared to the zero-shot method, aligning with past findings. All models surpass the baseline in detecting implicit hate speech, suggesting the features chosen are effective. Models that integrate linguistic features with additional completions from baseline prompts generally yield better outcomes than those using only linguistic features. However, they still do not match the performance of models trained with data from crowdworkers, indicating that GPT-3.5 cannot replace human annotators.

Figure 16 presents results from logistic regression models using data generated by GPT-4 for the same dataset and prompting approaches. In contrast to Figure 15, the few-shot approach in Figure 16 generally improves results over the zero-shot method, consistent with earlier experiments. The performance gap between models is narrower, showing GPT-4's enhanced robustness.

The best-performing models are Ling. Feat. + Completions DHSD and Ling. Feat. + Completion "offensive" closely followed by the model Ling. Feat., all three in the few-shot approach. The difference between Ling. Feat. and Ling. Feat. + Completion "offensive" is statistically significant. The t-statistic is -5.983, the p-value is 0.0039, and with a degree of freedom of 4, i.e., $\mathbf{n-1}$, where $\mathbf{n}$ is the number of cross-validation folds, which indicates a significant difference between the two conditions. This incremental improvement was assessed using a t-test on the F1 scores from cross-validation folds, employing the SciPy module for statistical functions called `scipy.stats` (Virtanen et al., 2020). The t-test results, showing a p-value under 0.05, confirm the statistical significance of these improvements. This indicates a significant advantage when linguistic features are

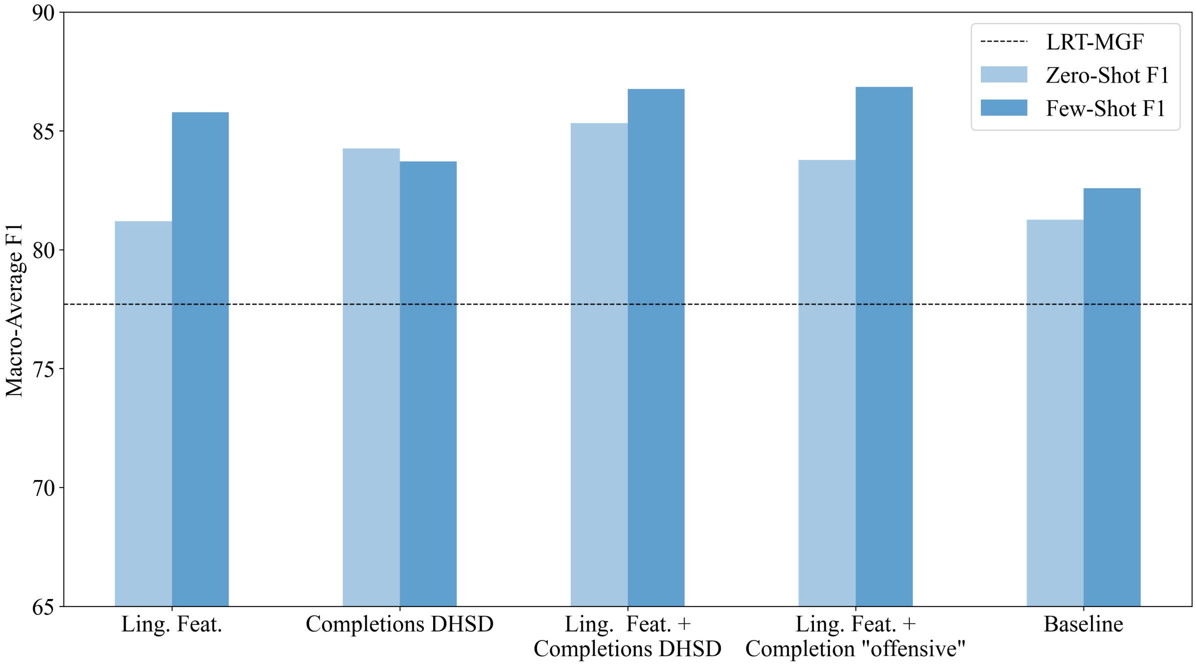


Figure 16.: Dataset *Identity Groups*: Logistic Regression Classifiers Trained on Data Generated by GPT-4

combined with completions from hate speech detection prompts.

The baseline model also performs well, especially when compared to its counterpart in the GPT-3.5 analysis. However, not as well as the logistic regression models. This indicates that using a logistic regression model trained on GPT-4 generated data is more effective at recognizing implicit hate speech than directly instructing GPT-4 to perform the classification task.

Furthermore, all models surpass the dotted line representing the performance of logistic regression trained on manually generated features, suggesting that GPT-4 can effectively substitute human annotators.

### 6.5.2. Dataset *Euphemistic Abuse*: Logistic Regression Trained on Data Generated by GPT Models

Figure 17 shows results from logistic regression models trained on the *Euphemistic Abuse* dataset with GPT-3.5 data, using both zero-shot and few-shot prompting. In this scenario, all models demonstrate lower F1 scores compared to other datasets, indicating that detecting euphemistic abuse is particularly challenging for GPT-3.5. The performances of the various models are quite similar, suggesting that neither adding completions nor integrating linguistic features significantly surpass the baseline. All models perform substantially below the dotted line representing logistic regression trained on manually generated features, reinforcing that GPT-3.5 cannot substitute for human annotators.

Figure 18 presents results from logistic regression models trained with GPT-4 data on the same *Euphemistic Abuse* dataset, again using both prompting approaches. As seen with previous datasets, all models trained with GPT-4 data show improved performance over those trained with data generated by GPT-3.5. Additionally, the baseline demonstrates

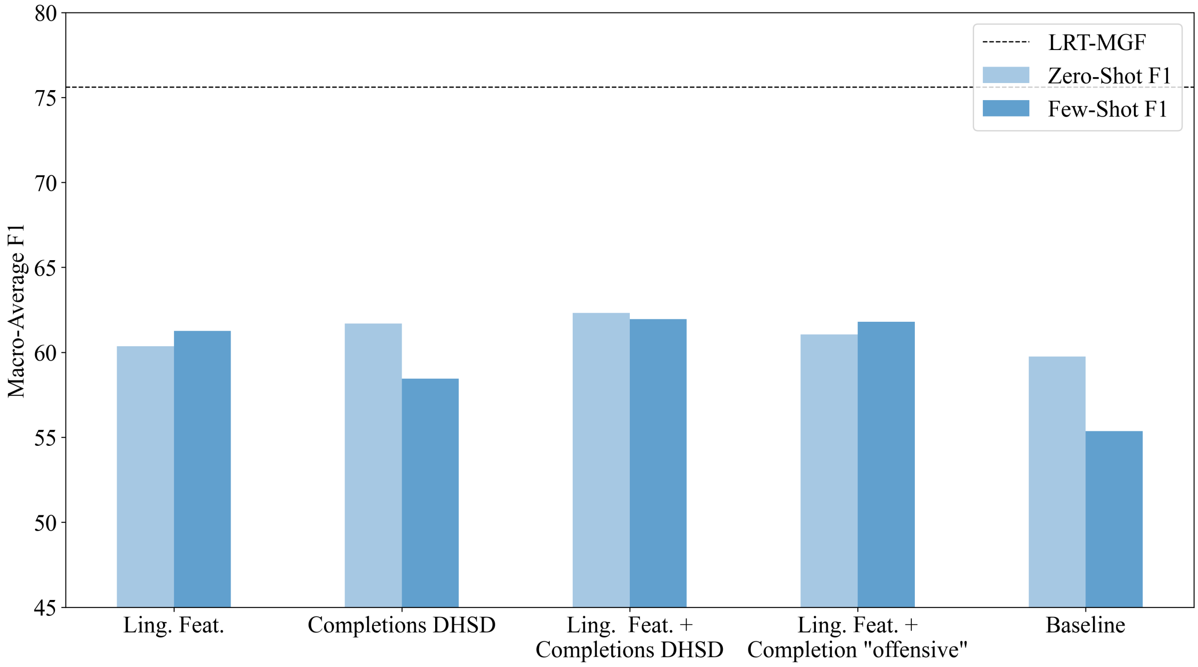


Figure 17.: Dataset *Euphemistic Abuse*: Logistic Regression Classifiers Trained on Data Generated by GPT-3.5

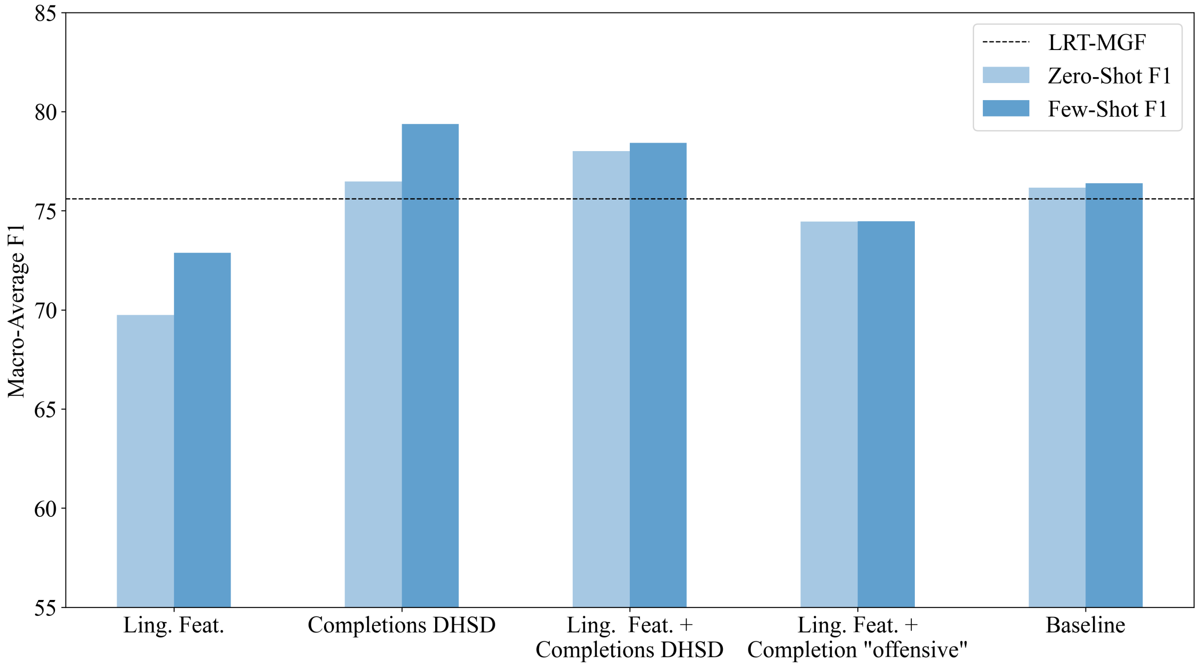


Figure 18.: Dataset *Euphemistic Abuse*: Logistic Regression Classifiers Trained on Data Generated by GPT-4

significant improvement over its GPT-3.5 counterpart.

Models combining linguistic features with completions, i.e., Ling. Feat. + Completions DHSD and Ling. Feat. + Completion “offensive”, perform well, with scores close to and even surpassing the performance represented by the LRT-MGF dotted line. This underscores the effectiveness of a multifaceted approach that uses linguistic insights alongside targeted prompt completions in identifying euphemistic abuse.

However, a t-test was conducted to compare the performance of the Baseline model and the Linguistic Features + Completions DHSD model. The t-statistic was 2.037, the p-value was 0.1113, with a degree of freedom of 4, i.e., $\mathbf{n}-\mathbf{1}$, where $\mathbf{n}$ is the number of cross-validation folds. The results showed that there is no statistically significant difference between the two models. Therefore, the difference in their performance is not large enough to be considered meaningful, and it could be due to random chance. The method that involves training a logistic regression model using data that was generated by GPT-4 is equally effective at identifying implicit forms of euphemistic abuse as the method that directly uses GPT-4 to classify sentences into abuse or other.

### 6.5.3. Dataset *Comparisons*: Logistic Regression Trained on Data Generated by GPT Models

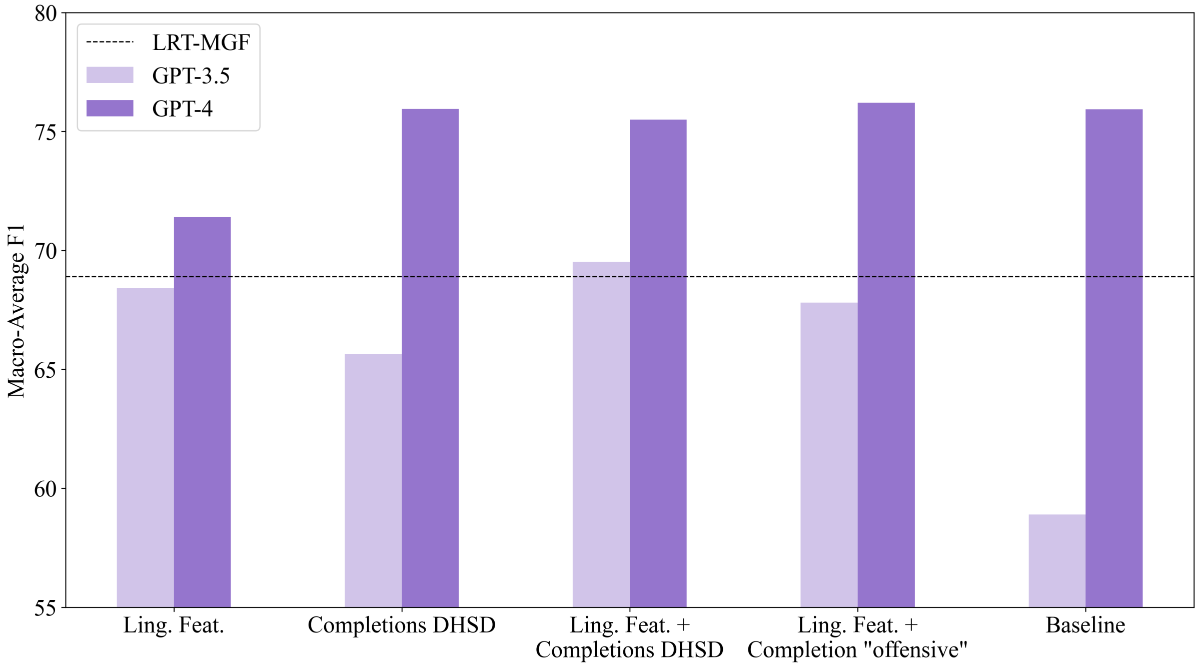


Figure 19.: Dataset *Comparisons*: Logistic Regression Classifiers Trained on Data Generated by GPT-3.5 and GPT-4 in the Zero-Shot Prompting Approach

Figure 19 shows the results of logistic regression models trained on the *Comparisons* dataset using data generated by GPT-3.5 and GPT-4 with zero-shot prompting. As discussed in Section 5.4.3, few-shot prompting for automated feature extraction was not entirely feasible. Therefore, Figure 19 compares the logistic regression results using zero-shot data from GPT-3.5 and GPT-4, rather than comparing zero-shot with few-shot prompting.

Most findings from previous figures are applicable here. GPT-4 demonstrates greater robustness and effectiveness in detecting implicit hate speech and extracting linguistic features. All models using GPT-4 data outperform the model based on manually generated data, with the Ling. Feat. + Completion “offensive” model achieving the highest F1 score. In contrast, the only model using GPT-3.5 data to exceed the performance of the manually generated data model is Ling. Feat. + Completions DHSD, and only by a narrow margin.

When using GPT-4 generated data, the model Ling. Feat. + Completions DHSD significantly outperforms the model Ling. Feat., as indicated by a t-test with a p-value of 0.05. This suggests that the completions from the direct hate speech detection with GPT-4 can be effectively used as features.

However, further analysis of the GPT-4 data revealed through t-tests that there is no statistically significant difference between the Baseline and the Ling. Feat. + Completion “offensive” model, nor between the Completions DHSD and Ling. Feat. + Completion DHSD models. The respective t-statistics were -0.204 and 0.872, with p-values of 0.848 and 0.432 and a degree of freedom of 4. This indicates that there is no measurable improvement over the baseline established in Section 6.2, and that these four models are equally effective at recognizing implicitly abusive comparisons. These findings are consistent with the ones from previous experiments with the dataset Euphemistic Abuse. Using a logistic regression model trained on GPT-4 generated data is just as effective as directly instructing GPT-4 to perform the classification task.

In conclusion, logistic regression models trained on data generated by GPT-4 consistently outperform those trained on data from GPT-3.5. Additionally, the few-shot prompting approach with GPT-4 improves results over the zero-shot method, whereas the few-shot approach with GPT-3.5 shows a reduction in performance compared to zero-shot. Furthermore, models using GPT-4 data generally surpass the performance of logistic regression models trained on manually generated features, suggesting that GPT-4 can effectively substitute human annotators in identifying implicit hate speech and its linguistic features. Moreover, models that combine linguistic features with completions from baseline prompts, such as Ling. Feat. + Completions DHSD, perform better than those using only linguistic features. Lastly, for the *Euphemistic Abuse* dataset, the difference in performance between the Baseline model and the Ling. Feat. + Completions DHSD model is not statistically significant, indicating similar effectiveness. Similar findings are observed in the *Comparisons* dataset. However, this was not the case for the *Identity Groups* dataset, where the logistic regression models trained on data generated by GPT-4 outperformed GPT-4’s capabilities at direct implicit hate speech detection. This indicates that using logistic regression trained on GPT-4 generated data is either equal to or better than directly instructing GPT-4 to perform the classification task.

# 7. Discussion

This chapter is organized into three sections. The first section outlines the major findings from the experiments. The second section discusses the limitations and challenges encountered during the study, including dataset constraints, temporal limitations, prompt variability, transparency and reproducibility in LLMs. The final section addresses ethical considerations, such as the implications of replacing human annotators with GPT models and the potential biases inherent in these models.

## 7.1. Major Findings

The performance of LLMs varied significantly across all investigated aspects, with the most recent and advanced model, GPT-4, demonstrating the strongest performance. This model proved to be effective in both extracting linguistic features and categorizing examples as implicit hate speech or non-hate speech.

The GPT-driven feature extraction showed varying results. Specifically, features such as non-episodic aspect, depiction of targets as perpetrators, use of taboo topics, contradictions, figurative language, and dehumanization were successfully extracted by the LLMs. Conversely, the LLMs struggled to recognize features like the depiction of targets as non-conformists, use of negated antonyms of abusive words, opposing sentiments, evaluative statements, and absurd images.

An important insight from the study is that few-shot prompting did not consistently improve results across all models. For instance, the performance of GPT-3.5 decreased with few-shot prompting compared to zero-shot results, indicating that this approach might not necessarily be beneficial depending on the model used. In contrast, GPT-4 showed improvement with few-shot learning, demonstrating its capability to refine its predictions with additional examples.

Furthermore, the proposed linguistic features were validated as crucial markers for detecting implicit hate speech. Training rule-based classifiers and logistic regression models on the features extracted by GPT-4 performed equally well or better than directly instructing GPT-4 to perform the classification task. This finding underscores the potential of these features as reliable indicators for discriminating implicit hate speech from non-hate speech.

Lastly, logistic regression models and rule-based classifiers using GPT-4 generated data showed superior results compared to models trained on manually generated features. This indicates that LLMs like GPT-4 can be effectively utilized to automate the extraction of these linguistic features, reducing the dependency on manual annotation.

These findings have notable implications for the development of automated systems for hate speech detection. They demonstrate that advanced LLMs, particularly GPT-4, can enhance the efficiency of implicit hate speech detection through effective feature extraction. This advancement not only streamlines the process but also ensures that subtle and covert forms of hate speech are more accurately identified, contributing to a safer and more respectful online environment.

## 7.2. Limitations and Challenges

This thesis acknowledges several limitations and challenges that impact the findings and conclusions. First, the datasets utilized in this thesis have notable constraints. As indicated by Wiegand et al. (2022), the *Identity Groups* dataset may not fully represent real-world implicit hate speech, potentially lacking an accurate reflection of actual occurrences. Composed solely of atomic negative sentences, this dataset represents single ideas and conveys negative sentiments. However, real-world hate speech is often more nuanced and does not always follow this straightforward pattern. For instance, consider the abusive sentence:

(1) "Once again, we find Jews and money money money." (Wiegand et al., 2022, 5608)

Unlike direct negative statements, this sentence subtly links a stereotype, i.e., Jews and an obsession with money, without explicit negative action by the target. Thus, hate speech can be veiled in statements that do not explicitly use negative verbs or direct insults.

Moreover, abuse does not always remain within sentence boundaries, and context might be necessary for its recognition. Wiegand et al. (2022) highlight the challenges of classifying certain sentences as factual statements or opinions influenced by stereotypical views. For example (Wiegand et al., 2022, 5608):

(2) "Women overuse makeup."

(3) "Muslims suppress Christian life in Iraq."

The interpretation of these sentences as abusive depends on perceiving them as opinions rather than facts. Additionally, the perception of a sentence's abusiveness can vary based on the reader's ideology or prior sentiment. Consider Examples (4) and (5) taken from Wiegand et al. (2022, 5608).

(4) "Muslims surrender to God's will."

(5) "Women unmake patriarchy."

Atheists might view Example (4) as abusive, while religious individuals may not. Similarly, feminists and non-feminists may perceive Example (5) differently. Accurately resolving these ambiguities is challenging without additional context.

The *Euphemistic Abuse* dataset has limitations as well due to its creation through crowdsourcing rather than extracting text from existing datasets or the Web, raising concerns about its authenticity. However, Wiegand et al. (2023) argue that, similarly to many other sub-types of implicit abuse, current datasets lack a sufficiently representative set of instances for these long-tail phenomena. This crowdsourcing strategy has also been employed for other sub-types of implicitly abusive language (Vidgen et al., 2021; Wiegand et al., 2021a) and fields such as plagiarism detection (Potthast et al., 2010) and deception detection (Ott et al., 2011).

Similarly, the *Comparisons* dataset was created by having crowdworkers invent instances of abusive language, potentially resulting in unauthentic data that may not fully capture natural language use in real-world scenarios. Wiegand et al. (2021a) acknowledge that creating a dataset via crowdsourcing inevitably results in artificial data. However, they argue that this method is necessary to produce a dataset of reasonable size with low bias, crucial for researching abusive comparisons.

Second, this study represents a snapshot in time. As new GPT models are continuously released, the findings and conclusions drawn here are likely to become outdated quickly. For instance, Llama-3 (AI@Meta, 2024) and GPT-4o (OpenAI, 2024) were released in May 2024, highlighting the rapid pace of development in this field. Future advancements in GPT technology or entirely new approaches may yield different results, necessitating ongoing evaluation and adaptation.

Third, the vast potential for prompt formulations introduces a virtually infinite search space. Outcomes can vary significantly depending on how prompts are phrased. Ideally, a much larger variety of prompts would have been tested to explore this variability, but time constraints limited the scope of prompt experimentation in this thesis. Still, a wide range of features and prompt variations were tested and reported here, offering a comprehensive foundation for future research in implicit hate speech detection. Additionally, due to the time gap between experiments, slight variations in the instructions of the prompts have occurred. Although small due to the high similarity, such as “No other answer is permitted” vs. “No other answer permitted,” these deviations could have possible effects on the predictions obtained.

Finally, according to Rogers (2023), closed LLMs should not be used to establish public benchmarks due to their lack of transparency and reproducibility. Without detailed knowledge of their architecture, training data, and methodologies, it is impossible to make meaningful comparisons or validate results. The unknown training data poses a risk of overlap between training and test datasets, leading to artificially improved outcomes. This potential contamination undermines the scientific integrity of research. Open, transparent models that can be independently verified and reproduced are essential for maintaining rigorous scientific standards in NLP research. However, this master’s thesis does not aim to establish a new public benchmark for implicit hate speech detection. The baseline described in Sections 5.3 and 6.2 was used merely for comparison within the proposed experiments. Based on testing and examples, an improvement with GPT-4 has been observed. But it is uncertain whether this improvement is due to an enhanced model architecture or proprietary data.

## 7.3. Ethical Considerations

The research conclusively demonstrates that GPT-4 can effectively replace human annotators for the tasks of implicit hate speech classification and detection of linguistic features of implicit hate speech, which are typically performed by crowdworkers (Mollas et al., 2020; Founta et al., 2018; Wiegand et al., 2022). Consequently, these findings could negatively impact the workforce of human annotators. Posch et al.’s (2022) findings suggest that for a significant portion of crowdworkers, especially those in economically disadvantaged regions such as Venezuela or India, microtask income is a crucial component of their financial stability. Similarly, Ross et al. (2010) discuss the evolution of the crowdworker population, noting a shift from a predominantly moderate-income, U.S.-based workforce to a more diverse international group, including a substantial number of young, well-educated Indian workers. This demographic change suggests that many workers may view microtasks as a full-time occupation, essential for their financial survival. If GPT models are used to replace these workers, it would significantly contribute to worldwide socio-economic inequality.

Nevertheless, considering the potential benefits of this development is important. Beyond economic impacts, the psychological and emotional burden of annotating hate speech should be addressed. Human annotators are often exposed to disturbing and harmful content, leading to psychological distress. Research indicates that content moderators, who regularly review such messages, experience significant emotional labor and are at risk of developing secondary traumatic stress, compassion fatigue, and burnout (Steiger et al., 2021). Employing LLMs to handle these tasks can alleviate the emotional and psychological strain on human annotators, providing relief and improving their overall well-being.

Moreover, Ding et al. (2022) argue that the democratization of deep learning seeks to make these technologies accessible across society, including to individuals, small and medium-sized enterprises, academic research labs, and nonprofit organizations. Achieving this objective could promote innovation, economic growth, fairness, and equality. However, one significant obstacle to deep learning democratization is the necessity of well-annotated data for training models, as they typically require large datasets. Ding et al. (2022) note that supervised learning heavily depends on sufficient training data with accurate annotations, and the data annotation process can be particularly costly for smaller companies and organizations. These costs encompass labor for data tagging, as well as the time and resources needed for hiring, training, and managing annotators. Additionally, there are expenses related to the required annotation tools and infrastructure. Ding et al. (2022) highlight that smaller entities might lack the resources to produce enough training data, thereby hindering their ability to benefit from advanced modeling techniques. While pre-trained language models such as BERT (Devlin et al., 2019), XLNet (Yang et al., 2019), and RoBERTa (Liu et al., 2019) help reduce some of the data demands, Ding et al. (2022) emphasize that data annotation continues to be a vital and unavoidable challenge in the training of supervised models. Substituting human annotators with GPT models might negatively impact the workforce, but it could also advance the democratization of deep learning.

Another major concern with GPT models is their tendency to perpetuate biases found in the datasets used for their training (Bender et al., 2021). These biases arise because GPT models are pre-trained on large volumes of unlabelled data, which can contain inherent biases and stereotypes. Bender et al. (2021) pointed out that the vast datasets used in training GPT models are assumed to represent diverse perspectives but are biased by the narrow scope of internet participation and filtering processes. Consequently, hegemonic viewpoints, including white supremacist, misogynistic, and ageist views are overrepresented, amplifying these biases in the models.

Wang et al. (2023) assessed the trustworthiness of GPT-3.5 and GPT-4, particularly their susceptibility to biases and adversarial attacks. They found that while the models generally avoid biased content, they can be manipulated by targeted prompts to produce biased outputs. For example, GPT-4 showed different levels of bias depending on the prompt and topic, displaying more biased content in response to certain stereotypes. However, when presented with benign system prompts, both GPT models generally reject biased statements for most stereotype-related topics.

Nevertheless, to tackle the issue of bias in GPT models, it is essential to ensure that the training data for GPT models is diverse and encompasses a variety of experiences and perspectives. This becomes problematic because GPT-3.5 and GPT-4 are proprietary, and OpenAI does not disclose their training data. Therefore, ongoing monitoring and evaluation of the models' outputs are crucial to detect and correct any potential biases.

# 8. Conclusion

This study investigated the effectiveness of LLMs in extracting linguistic features of implicit hate speech in English and their contribution to automated detection. The research primarily focused on evaluating LLaMA-2, GPT-3.5, and GPT-4.

LLaMA-2, an open-source alternative, faced technical issues and poor task comprehension, leading to its discontinuation for the task of linguistic feature extraction. GPT-3.5's performance was suboptimal, highlighting difficulties in recognizing implicitly abusive language and its linguistic features. In contrast, GPT-4 showed substantial promise, particularly when employing few-shot learning to extract complex linguistic features. The findings indicated that rule-based classifiers utilizing data generated by GPT-4 outperformed both GPT-4's direct detection capabilities and classifiers based on manually extracted features. Similarly, logistic regression models trained on GPT-4 generated data demonstrated superior results compared to those trained on manually generated features. This implies that leveraging linguistic features extracted by GPT-4 can significantly enhance the detection of implicit hate speech.

The relevance of this research lies in addressing the challenge of limited availability of accurately labeled data for implicit hate speech detection. By reducing the reliance on manually annotated data, the proposed approach offers a potential solution to this issue. The study's findings suggest that GPT-4 shows potential as a substitute for human annotators in identifying linguistic features of implicitly abusive language.

However, several limitations were identified. The proprietary nature of GPT models' training data raises concerns about transparency and potential biases. Moreover, the socio-economic implications of replacing human annotators, particularly in economically disadvantaged regions, warrant careful consideration.

Future work could explore several avenues to enhance LLMs' effectiveness in hate speech detection. Refining prompt formulations in few-shot and zero-shot scenarios could lead to further performance improvements. Additionally, applying similar approaches with more recent GPT models and extending research to other languages, particularly low-resource languages, could broaden the impact and utility of these findings. Collaborative research involving linguists and experts in various languages would be instrumental in investigating whether the linguistic features used as indicators of hate speech in English are valid across different languages. Moreover, GPT models could be explored as tools for annotating data for other NLP tasks. This would not only reduce the reliance on labor-intensive manual annotations but also improve the overall accuracy and efficiency of various NLP applications.

In conclusion, this thesis suggests that utilizing GPT-4 for extracting linguistic features and training classifiers can enhance the detection of implicit hate speech. The proposed hybrid approach combines GPT-driven feature extraction with supervised feature-based machine learning techniques, simplifying the typically complex process of feature extraction. These findings provide a promising tool for addressing this pervasive issue, underscoring the potential of advanced NLP techniques in improving automated hate speech detection.

# Bibliography

AI@Meta (2024). Llama 3 model card. Accessed: 2024-07-20.

Keith Allan and Kate Burridge (2006). *Taboos and their origins*, page 1–28. Cambridge University Press.

Ayme Arango, Jorge Pérez, and Barbara Poblete (2019). Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR)*, pages 45–53, Paris, France.

John Langshaw Austin (1975). *How To Do Things With Words: The William James Lectures delivered at Harvard University in 1955*. Oxford University Press.

C. Edwin Baker (2012). Hate speech. In Michael Herz and Peter Molnar, editors, *The Content and Context of Hate Speech: Rethinking Regulation and Responses*, page 57–80. Cambridge University Press.

Jawid Ahmad Baktash and Mursal Dawodi (2023). GPT-4: A Review on Advancements and Opportunities in Natural Language Processing. *ArXiv*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti (2019a). SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti (2019b). SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pages 610–623, Virtual Event, Canada. Association for Computing Machinery.

Whorf Benjamin Lee, Carroll John B., Levinson Stephen C., and Lee Penny (2012). *Language, Thought, and Reality : Selected Writings of Benjamin Lee Whorf.*, volume 2nd ed of *The Massachusetts Institute of Technology Press.* The Massachusetts Institute of Technology Press.

Lorraine Bowman-Grieve (2009). Exploring "stormfront": A virtual community of the radical right. *Studies in Conflict & Terrorism*, 32(11):989–1007.

Alexander Brown (2017a). What is hate speech? part 1: The myth of hate. *Law and Philosophy*, 36(4):419–468.

Alexander Brown (2017b). What is hate speech? part 2: Family resemblances. *Law and philosophy*, 36(5):561–613.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020a). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020b). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Pete Burnap and Matthew L. Williams (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer (2020). I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Xuanting Chen, Junjie Ye, Can Zu, Nuo Xu, Rui Zheng, Minlong Peng, Jie Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang (2023). How robust is gpt-3.5 to predecessors? a comprehensive study on language understanding tasks. *ArXiv*, abs/2303.00293.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini (2019). CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.

Constitutional Assembly of South Africa (1996). Chapter 2: Bill of Rights, Constitution of the Republic of South Africa. Available: `https://www.justice.gov.za/legislation/constitution/SAConstitution-web-eng-02.pdf`.

Anthony Joseph Paul Cortese (2006). *Opposing hate speech.* Praeger Publishers, Westport, Connecticut.

Matthew Costello and James Hawdon (2018). Who are the online extremists among us? sociodemographic characteristics, social networking, and online experiences of those who produce online hate materials. *Violence and Gender*, 5(1):55–60.

Council of Europe, Committee of Ministers (1997). Recommendation No. R (97) 20 of the Committee of Ministers to Member States on "Hate Speech". *Council of Europe Official Documents*. Adopted at the 607th meeting of the Ministers' Deputies.

Kevin Crowston (2012). Amazon mechanical turk: A research tool for organizations and information systems scholars. In *Shaping the Future of ICT Research. Methods and Approaches*, pages 210–221, Berlin, Heidelberg. Springer Berlin Heidelberg.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*.

Phillip Davis (1996). Threats of corporal punishment as verbal aggression: A naturalistic study. *Child Abuse & Neglect*, 20(4):289–304.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros (2018). Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Richard Delgado (1982). Words that wound: A tort action for racial insults, epithets, and name-calling. *Harvard civil rights-civil liberties law review*, 17(1):133.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq R. Joty, and Boyang Albert Li (2022). Is gpt-3 a good data annotator? In *Annual Meeting of the Association for Computational Linguistics*.

Ronald Dworkin (2009). Foreword. In Ivan Hare and James Weinstein, editors, *Extreme Speech and Democracy*. Oxford University Press.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang (2021). Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

European Court of Human Rights (1999). Case of sürek v. turkey (no. 1). Judgment, Strasbourg. Application no. 26682/95.

European Court of Human Rights (2003). Case of gündüz v. turkey. Judgment, Strasbourg. Application no. 35071/97.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini (2021). Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240, Online. Association for Computational Linguistics.

Lucie Flekova and Iryna Gurevych (2016). Supersense embeddings: A unified model for supersense interpretation, prediction, and utilization. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2029–2041, Berlin, Germany. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes (2018). A survey on automatic detection of hate speech in text. *Association for Computing Machinery Computing Surveys*, 51(4).

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12.

Annemarie Friedrich and Manfred Pinkal (2015). Automatic recognition of habituals: a three-way classification of clausal aspect. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2471–2481, Lisbon, Portugal. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou (2020). Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Katharine Gelber and Luke McNamara (2016). Evidencing the harms of hate speech. *Social identities*, 22(3):324–341.

David Graff (1995). North american news text corpus. Web Download. LDC Catalog No.: LDC95T21. ISBN: 1-58563-053-5. ISLRN: 667-148-284-023-7. Member Years: 1995, 1996, 1997. DCMI Type(s): Text. Data Source(s): newswire. Projects: TIDES, MUC, Hub4, GALE, EARS. Applications: language modeling, information retrieval. Languages: English. Language ID(s): eng. License(s): North American News Text Agreement. Online Documentation: LDC95T21 Documents. Licensing Instructions: Subscription & Standard Members, and Non-Members.

Hugo Lewi Hammer (2017). Automatic detection of hateful comments in online discussion. In *Industrial Networks and Intelligent Systems*, pages 164–173, Cham. Springer International Publishing.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar (2022). ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

HateAid (2022). Suizid von Ärztin kellermayr: Was hass im netz anrichtet. Accessed on December 12, 2023.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen (2021). Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Joe Hoover (2023). Run llama 2 with an api. `https://replicate.com/blog/run-llama-2-with-an-api?input=python`. Accessed: 2024-04-25.

Fan Huang, Haewoon Kwak, and Jisun An (2023). Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 294–297, New York, NY, USA. Association for Computing Machinery.

Susan Hurley (2004). Imitation, media violence, and freedom of speech. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 117(1/2):165–218.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin (2024a). Fine-tuning and masked language models. In *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, chapter 11. Pearson Prentice Hall, Upper Saddle River, N.J.

Daniel Jurafsky and James H. Martin (2024b). Logistic regression. In *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, chapter 5. Pearson Prentice Hall, Upper Saddle River, N.J.

Daniel Jurafsky and James H. Martin (2024c). Naive bayes, text classification, and sentiment. In *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, chapter 4. Pearson Prentice Hall, Upper Saddle River, N.J.

Daniel Jurafsky and James H. Martin (2024d). Neural networks and neural language models. In *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, chapter 7. Pearson Prentice Hall, Upper Saddle River, N.J.

Daniel Jurafsky and James H. Martin (2024e). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd edition. Pearson Prentice Hall, Upper Saddle River, N.J.

Daniel Jurafsky and James H. Martin (2024f). Transformers and large language models. In *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, chapter 10. Pearson Prentice Hall, Upper Saddle River, N.J.

T. Keipi, M. Näsi, A. Oksanen, and P. Räsänen (2016). *Online Hate and Harmful Content: Cross-National Perspectives*. Routledge Advances in Sociology. Taylor & Francis.

B. Kennedy, M. Atari, A. M. Davani, et al. (2022). Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56:79–108.

Chris Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano (2020). Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Natalia Knoblock (2022). Introduction. In Natalia Knoblock, editor, *The Grammar of Hate: Morphosyntactic Features of Hateful, Aggressive, and Dehumanizing Discourse*, pages 1–14. Cambridge University Press.

J. Richard Landis and Gary G. Koch (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Rae Langton (1993). Speech acts and unspeakable acts. *Philosophy & Public Affairs*, 22(4):293–330.

Charles R. Lawrence (1990). If he hollers let him go: Regulating racist speech on campus. *Duke law journal*, 1990(3):431–483.

Michiel Leezenberg (2015). Discursive violence and responsibility: Notes on the pragmatics of dutch populism. *Journal of Language Aggression and Conflict*, 3(1):200–228.

Cecilia H. Leonard, George D. Annas, James L. Knoll, and Terje Torrissen (2014). The Case of Anders Behring Breivik: Language of a Lone Terrorist. *Behavioral Sciences & the Law*, 32:408–422.

Hongzhan Lin, Ziyang Luo, Bo Wang, Ruichao Yang, and Jing Ma (2024). Goat-bench: Safety insights to large multimodal models through meme-based social abuse.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Catharine Alice MacKinnon (1993). *Only Words*. Harvard University Press, Cambridge, Massachusetts.

Ishani Maitra and Mary Kate McGowan (2012). *Speech and Harm: Controversies Over Free Speech*, 1 edition. Oxford University Press, Oxford.

Kate Malleson (2018). Equality law and the protected characteristics. *Modern law review*, 81(4):598–621.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee (2021). Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Puneet Mathur, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah (2018). Did you offend me? classification of offensive tweets in Hinglish language. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 138–148, Brussels, Belgium. Association for Computational Linguistics.

Mari J. Matsuda (1989). Public response to racist speech: Considering the victim's story. *Michigan law review*, 87(8):2320–2381.

Julia Mendelsohn, Ceren Budak, and David Jurgens (2021). Modeling framing in immigration discourse on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263, Online. Association for Computational Linguistics.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky (2020). A framework for the computational linguistic analysis of dehumanization. *Frontiers in Artificial Intelligence*, 3.

Meta Platforms, Inc. (2023). Community standards on hate speech. `https://transparency.meta.com/policies/community-standards/hate-speech/`. Accessed: 2023-04-18.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller (1990). Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4):235–244.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi (2022a). MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer (2022b). Rethinking the role of demonstrations: What makes in-context learning work? In *Empirical Methods in Natural Language Processing*.

Saif Mohammad and Peter Turney (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas (2020). ETHOS: an online hate speech detection dataset. *CoRR*, abs/2006.08328.

Brian Mullen and Joshua M. Smyth (2004). Immigrant suicide rates as a function of ethnophaulisms: Hate speech predicts death. *Psychosomatic Medicine*, 66:343–348.

Karsten Müller and Carlo Schwarz (2017). Fanning the flames of hate: Social media and hate crime. *SSRN Electronic Journal*.

Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata (2023). 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*, pages 1997–2013, Dubrovnik, Croatia. Association for Computational Linguistics.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph (2023). Gpt-4 technical report.

OpenAI (2024). Hello, gpt-4o. Accessed: 2024-06-06.

OpenAI Playground (2024). Playground. https://platform.openai.com/playground. Accessed: 2024-04-25.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319, Portland, Oregon, USA. Association for Computational Linguistics.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung (2019). Multilingual and multi-aspect hate speech analysis. 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)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Oxford English Dictionary (2023). perpetrator, n. Accessed on December 11, 2023.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos (2020). Toxicity detection: Does context really matter? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4296–4305, Online.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy (2023). Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada. Association for Computational Linguistics.

Lisa Posch, Arnim Bleier, Fabian Flöck, Clemens Lechner, Katharina Kinder-Kurlanda, Denis Helic, and Markus Strohmaier (2022). Characterizing the global crowd workforce: A cross-country comparison of crowdworker demographics. *Human Computation*, 9(1):22–57.

Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso (2010). An evaluation framework for plagiarism detection. In *Proceedings of the International*

*Conference on Computational Linguistics (COLING)*, pages 997–1005, Beijing, China. Coling 2010 Organizing Committee.

Prolific (2023). Prolific. `https://www.prolific.com`. Accessed: 2023-04-18.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1).

Santhosh Rajamanickam, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova (2020). Joint modelling of emotion and abusive language detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4270–4279, Online.

Nils Reimers and Iryna Gurevych (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Republic of South Africa (2000). Promotion of Equality and Prevention of Unfair Discrimination Act 4 of 2000. Date of Commencement: 16 June 2003.

Thorsten Roelcke (2010). *Fachsprachen*, 3., neu bearbeitete auflage edition. Grundlagen der Germanistik ; 37. E. Schmidt, Berlin.

Anna Rogers (2023). Closed ai models make bad baselines. *Hacking Semantics.*

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki (2016). Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *ArXiv*, abs/1701.08118.

Joel Ross, Lilly Irani, Six Silberman, Andrew Zaldivar, and Bill Tomlinson (2010). Who are the crowdworkers? shifting demographics in mechanical turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, page 2863–2872, New York, NY, USA. Association for Computing Machinery.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah Smith, and Yejin Choi (2020). Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Julie Schumacher and Kenneth Leonard (2005). Husbands' and wives' marital adjustment, verbal aggression, and physical aggression as longitudinal predictors of physical aggression in early marriage. *Journal of consulting and clinical psychology*, 73(1):28–37.

Danny Scoccia (1996). Can liberals support a ban on violent pornography? *Ethics*, 106(4):776–799.

Leandro Silva, Mainack Mondal, Denzil Correa, Fabricio Benevenuto, and Ingmar Weber (2016). Analyzing the targets of hate in online social media.

Miriah Steiger, Timir Bharucha, Sukrit Venkatagiri, Martin Riedl, and Matthew Lease (2021). The psychological well-being of content moderators: The emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

J.E Stets (1990). Verbal and physical aggression in marriage. *Journal of marriage and family*, 52(2):501–514.

Malliga Subramanian, Veerappampalayam Easwaramoorthy Sathiskumar, G. Deepalakshmi, Jaehyuk Cho, and G. Manikandan (2023). A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. *Alexandria Engineering Journal*, 80:110–121.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar (2020). Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Henri Tajfel and John Turner (1979). An integrative theory of intergroup conflict. In W. G. Austin and S. Worchel, editors, *The social psychology of intergroup relations*, pages 33–47. Brooks/Cole, Monterey, CA.

The Pandas Development Team (2024). pandas-dev/pandas: Pandas.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom (2023). Llama 2: Open foundation and fine-tuned chat models.

United Nations Human Rights: Office of the High Commissioner (2019). Joint open letter on concerns about the global increase in hate speech. `https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25036&LangID=E`. Accessed: 11.04.2024.

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser (2018). Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, Brussels, Belgium. Association for Computational Linguistics.

Cynthia Van Hee, Els Lefever, and Véronique Hoste (2018). SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Guido Van Rossum and Fred Drake (2009). *Python 3 Reference Manual.* CreateSpace, Scotts Valley, CA.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela (2021). Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Denny Vrandečić and Markus Krötzsch (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Jeremy Waldron (2010). Dignity and defamation: The visibility of hate. *Harvard Law Review*, 123(7):1596–1657.

Jeremy Waldron (2012). *The Harm in Hate Speech.* The Oliver Wendell Holmes Lectures; 2009. Harvard University Press,, Cambridge, Mass.:.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. (2023). Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track.*

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber (2017). Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy (2016a). Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy (2016b). Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Caroline West (2012). Words That Silence? Freedom of Expression and Racist Hate Speech. In *Speech and Harm: Controversies Over Free Speech*. Oxford University Press.

Michael Wiegand, Elisabeth Eder, and Josef Ruppenhofer (2022). Identifying implicitly abusive remarks about identity groups using a linguistically informed approach. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5600–5612, Seattle, United States. Association for Computational Linguistics.

Michael Wiegand, Maja Geulig, and Josef Ruppenhofer (2021a). Implicitly abusive comparisons – a new dataset and linguistic analysis. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 358–368, Online. Association for Computational Linguistics.

Michael Wiegand, Jana Kampfmeier, Elisabeth Eder, and Josef Ruppenhofer (2023). Euphemistic abuse – a new dataset and classification experiments for implicitly abusive language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16280–16297, Singapore. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder (2021b). Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer (2019). Detection of abusive language: The problem of biased datasets. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 602–608, Minneapolis, MN, USA.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg (2018). Inducing a lexicon of abusive words – a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056, New Orleans, Louisiana. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc Le (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*, pages 5753–5763, Red Hook, NY, USA. Curran Associates Inc.

Xinyi Zhou and Reza Zafarani (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *Association for Computing Machinery Computing Surveys*, 53(5).

# A. Appendix

## A.1. Dataset *Comparisons*: Implicit Hate Speech Detection with GPT-4 using Zero-Shot and Few-Shot Prompting

Table 18.: Dataset *Comparisons*: Evaluation Metrics for Implicit Hate Speech Detection with GPT-4 using Few-Shot Prompting

| **Marker** | **P. Neg.** | **R. Neg.** | **P. Pos.** | **R. Pos.** | **Av. P.** | **Av. R.** | **Av. F1** |
|---|---:|---:|---:|---:|---:|---:|---:|
| Abusive | 86.40 | **47.00** | 63.60 | 92.60 | 75.00 | 69.80 | 72.31 |
| Hateful | 85.88 | **43.80** | 62.28 | 92.80 | 74.08 | 68.30 | 71.07 |
| Insulting | 90.56 | **32.60** | 58.90 | 96.60 | 74.73 | 64.60 | 69.30 |
| Offensive | 87.72 | **40.00** | 61.14 | 94.40 | 74.43 | 67.20 | 70.63 |
| Toxic | 83.33 | **48.00** | 63.48 | 90.40 | 73.41 | 69.20 | 71.24 |

As discussed in Section 5.6, the baseline for implicit hate speech detection with GPT models was evaluated as follows. One linguistic marker of hate speech, i.e., abusive, hateful, insulting, offensive, or toxic, was identified as the positive class, and the prediction other was identified as the negative class for binary classification. Precision and recall were calculated for both classes and averaged to get a single precision and recall value. These values were used to determine the average macro F1 score. This method matched the evaluation approach used by Wiegand et al. (2021a, 2022, 2023), allowing for direct and reliable comparisons with their findings.

In contrast to GPT-3.5, GPT-4 demonstrated a strong ability to recognize implicit hate speech in the zero-shot approach, and showed further improvement with few-shot learning, as seen in Table 12. The exception was with the *Comparisons* dataset where the F1 score decreased. A closer inspection of recall and accuracy revealed many false negatives in the class other, i.e., the negative class. This issue is highlighted by the low recall for the negative class, as detailed in Table 18. A subsequent attempt with few-shot learning, providing 15 examples to the GPT model, five for the positive class and ten for the negative, resulted in improvements, as shown in Table 13. The corresponding prompt template can be seen here:

> Sentence: “With your weight you have to take 3 seats at the cinema.” Answer: “YES”
>
> Sentence: “You look like death warmed up.” Answer: “NO”
>
> Sentence: “You look like you have a case of the flu.” Answer: “NO”
>
> Sentence: “A snail could do this job faster than you.” Answer: “YES”
>
> Sentence: “You are as pale as a ghost.” Answer: “NO”

Sentence: “Your expression is like your pet has died.” Answer: “NO”

Sentence: “Your drawings have the charm of a car crash.” Answer: “YES”

Sentence: “Your expression reminds me of a thunderstorm.” Answer: “NO”

Sentence: “You are as tall as a tree.” Answer: “NO”

Sentence: “Your mouth reminds me of a fish.” Answer: “YES”

Sentence: “Your eyes are as hard as granite.” Answer: “NO”

Sentence: “Your progress is as slow as an iceberg.” Answer: “NO”

Sentence: “You eat like a pig.” Answer: “YES”

Sentence: “Your car is as green as an olive.” Answer: “NO”

Sentence: “You talk like someone with lots on their mind.” Answer: “NO”

Is this sentence {MARKER}? “{SENTENCE}” Answer either “YES” or “NO”.
No other answer permitted.

## A.2. Dataset *Identity Groups*: List of Prompts Used for GPT-Driven Feature Extraction

For the sake of completeness, Table 19 reports on the results of the prompts listed in Table 20. Table 20 lists all the prompts tested for the GPT-driven extraction of linguistic features of implicit hate speech present in the dataset *Identity Groups*. The relevant experiments are described in Section 5.4.1.

Table 19.: Dataset *Identity Groups*: All Results for the Zero-Shot Automated Extraction of Linguistic Feature of Implicit Hate Speech with GPT Models

| **Feature** | **Prompt from Table 20** | **Model** | **F1** |
|---|---|---|---|
| Non-Conformist | Prompt 1 | GPT-3.5 | 51.06 |
| Non-Conformist | Prompt 2 | GPT-3.5 | **58.97** |
| Aspect | Prompt 3 | GPT-3.5 | 65.50 |
| Aspect | Prompt 4 | GPT-3.5 | 63.81 |
| Aspect | Prompt 5 | GPT-3.5 | 73.28 |
| Aspect | Prompt 6 | GPT-3.5 | **73.95** |
| Perpetrator | Prompt 7 | GPT-3.5 | **69.31** |
| Perpetrator | Prompt 8 | GPT-3.5 | 63.50 |
| Non-Conformist | Prompt 1 | GPT-4 | 53.08 |
| Non-Conformist | Prompt 2 | GPT-4 | **62.58** |
| Aspect | Prompt 3 | GPT-4 | 82.66 |
| Aspect | Prompt 4 | GPT-4 | 82.28 |
| Aspect | Prompt 5 | GPT-4 | **83.49** |
| Aspect | Prompt 6 | GPT-4 | 83.06 |
| Perpetrator | Prompt 7 | GPT-4 | **82.24** |
| Perpetrator | Prompt 8 | GPT-4 | 79.90 |

Table 20.: Dataset *Identity Groups*: List of Prompts Used for GPT-Driven Extraction of Linguistic Features of Implicit Hate Speech

| **Prompt** | **Text** |
|---|---|
| Prompt 1 | Non-conformist views are sentences in which the sentiment of the person performing the action (agent) towards the person or the thing receiving the action (patient) disagrees with the sentiment of the patient. In this context, consider this sentence “{SENTENCE}” Does the author of this sentence think that {TARGET} are non-conformist? Answer either YES or NO. No other answer permitted. |
| Prompt 2 | Consider this sentence “{SENTENCE}” Does the author of this sentence have a negative sentiment towards {TARGET}? Answer either YES or NO. No other answer permitted. |
| Prompt 3 | An episodic aspect is an utterance that expresses information about a single event. Non-episodic means habitual, routine or regularly occurring. Given these definitions, is the following sentence episodic? “{SENTENCE}” Answer either YES or NO. No other answer permitted. |
| Prompt 4 | Episodic means expressing information about a singular, distinctive or particular event. Non-episodic means expressing information about a reoccurring event, about something resorted to on a regular basis, something of a commonplace character, typical, normal, or standard. Given these definitions, is the following sentence episodic? “{SENTENCE}” Answer either YES or NO. No other answer permitted. |
| Prompt 5 | Classify the following sentence as either “EPISODIC” or “NON-EPISODIC”: “{SENTENCE}” No other answer permitted. Answer “EPISODIC” if the sentence expresses information about a single event. Answer “NON-EPISODIC” if the sentence expresses information about something habitual, routine, or regularly occurring. |
| Prompt 6 | No other answer permitted. Answer “EPISODIC” if the sentence expresses information about a singular, distinctive or particular event. Answer “NON-EPISODIC” if the sentence expresses information about a reoccurring event, about something resorted to on a regular basis, something of a commonplace character, typical, normal, or standard. |
| Prompt 7 | A perpetrator is a person who commits an illegal, criminal, harmful or evil act. Consider the following sentence: “{SENTENCE}” Does the author of this sentence think that {TARGET} are perpetrators? Answer either YES or NO. No other answer permitted. |
| Prompt 8 | A wrongdoer is a person who engages in unlawful, illicit, detrimental, or malevolent actions. Consider the following sentence: “{SENTENCE}” Does the author of this sentence think that {TARGET} are wrongdoers? Answer either YES or NO. No other answer permitted. |

## A.3. Implicit Hate Speech Detection with Rule-Based Classifiers

Figure 14 illustrates the progress made with the rule-based classifier. Table 21 delivers the values corresponding to the bars in Figure 14.

Table 21.: Implicit Hate Speech Detection with Rule-Based Classifiers

| Classifier | Model | Zero-Shot F1 | Few-Shot F1 |
|---|---|---:|---:|
| Rule-Based | GPT-3.5 | 72.70 | 59.60 |
| Baseline | GPT-3.5 | 64.91 | 55.70 |
| Rule-Based | GPT-4 | 82.75 | 86.20 |
| Baseline | GPT-4 | 81.26 | 82.59 |
| Linguistically Informed | Wiegand et al. (2022, 5607) | 77.70 | 77.70 |

## A.4. Logistic Regression Models Across Various Datasets and Prompting Approaches

The goal of the experiments described in Section 5.5 was to determine if the features extracted automatically by GPT-3.5 and GPT-4 were predictive of implicitly abusive language. To this end, different logistic regression models were trained. Table 22 shows the F1 scores for each model. The values presented here corresponds to the values used in Figures 15, 16, 17, 18, and 19.

Table 22.: Logistic Regression Models Across Various Datasets and Prompting Approaches

| Dataset | Logistic Regression Trained on | Zero-Shot F1 | Few-Shot F1 | Model | Figure |
|---|---|---|---|---|---|
| Identity Groups | Ling. Feat. | 72.08 | 62.47 | GPT-3.5 | Figure 15 |
| Identity Groups | Completions DHSD | 66.89 | 63.83 | GPT-3.5 | Figure 15 |
| Identity Groups | Ling. Feat. + Completions DHSD | 74.53 | 65.27 | GPT-3.5 | Figure 15 |
| Identity Groups | Ling. Feat. + Completion “offensive” | 74.51 | 57.89 | GPT-3.5 | Figure 15 |
| Identity Groups | Baseline | 64.91 | 55.70 | GPT-3.5 | Figure 15 |
| Identity Groups | Ling. Feat. | 81.20 | 85.78 | GPT-4 | Figure 16 |
| Identity Groups | Completions DHSD | 84.26 | 83.71 | GPT-4 | Figure 16 |
| Identity Groups | Ling. Feat. + Completions DHSD | 85.33 | 86.76 | GPT-4 | Figure 16 |
| Identity Groups | Ling. Feat. + Completion “offensive” | 83.78 | 86.85 | GPT-4 | Figure 16 |
| Identity Groups | Baseline | 81.26 | 82.59 | GPT-4 | Figure 16 |
| Comparisons | Ling. Feat. | 68.41 |  | GPT-3.5 | Figure 19 |
| Comparisons | Completions DHSD | 65.65 |  | GPT-3.5 | Figure 19 |
| Comparisons | Ling. Feat. + Completions DHSD | 69.52 |  | GPT-3.5 | Figure 19 |
| Comparisons | Ling. Feat. + Completion “offensive” | 67.80 |  | GPT-3.5 | Figure 19 |
| Comparisons | Baseline | 58.90 |  | GPT-3.5 | Figure 19 |
| Comparisons | Ling. Feat. | 71.40 |  | GPT-4 | Figure 19 |
| Comparisons | Completions DHSD | 75.95 |  | GPT-4 | Figure 19 |
| Comparisons | Ling. Feat. + Completions DHSD | 75.50 |  | GPT-4 | Figure 19 |
| Comparisons | Ling. Feat. + Completion “offensive” | 76.20 |  | GPT-4 | Figure 19 |
| Comparisons | Baseline | 75.93 |  | GPT-4 | Figure 19 |
| Euphemistic Abuse | Ling. Feat. | 60.36 | 61.27 | GPT-3.5 | Figure 17 |
| Euphemistic Abuse | Completions DHSD | 61.69 | 58.45 | GPT-3.5 | Figure 17 |
| Euphemistic Abuse | Ling. Feat. + Completions DHSD | 62.32 | 61.95 | GPT-3.5 | Figure 17 |
| Euphemistic Abuse | Ling. Feat. + Completion “offensive” | 61.05 | 61.80 | GPT-3.5 | Figure 17 |
| Euphemistic Abuse | Baseline | 59.76 | 55.37 | GPT-3.5 | Figure 17 |
| Euphemistic Abuse | Ling. Feat. | 69.74 | 72.88 | GPT-4 | Figure 18 |
| Euphemistic Abuse | Completions DHSD | 76.47 | 79.38 | GPT-4 | Figure 18 |
| Euphemistic Abuse | Ling. Feat. + Completions DHSD | 78.01 | 78.43 | GPT-4 | Figure 18 |
| Euphemistic Abuse | Ling. Feat. + Completion “offensive” | 74.45 | 74.47 | GPT-4 | Figure 18 |
| Euphemistic Abuse | Baseline | 76.16 | 76.38 | GPT-4 | Figure 18 |

## A.5. *Identity Groups*, *Euphemistic Abuse* and *Comparisons*: GPT-Driven Hate Speech Detection

In order to assess how well the GPT models could identify implicit hate speech, zero-shot and few-shot approaches were devised. Each prompt followed the format:

> markers = [“hateful”, “abusive”, “offensive”, “toxic”, “insulting”]
>
> Is this sentence MARKER? “SENTENCE” Answer either “YES” or “NO”. No other answer is permitted.

In the few-shot approach the prompt was paired with ten examples, five examples representing each class. Table 23 reports on the results.

Table 23.: Datasets *Identity Groups, Euphemistic Abuse* and *Comparisons*: GPT-Driven Implicit Hate Speech Detection with Zero-Shot and Few-Shot Prompting

| | | | **Macro-Average F1** | **Macro-Average F1** |
|---|---|---|---:|---:|
| **Dataset** | **Model** | **Marker** | **Zero-Shot** | **Few-Shot** |
| Identity Groups | LLaMA-2 | Hateful | 54.72 | |
| Identity Groups | LLaMA-2 | Insulting | **55.97** | |
| Identity Groups | LLaMA-2 | Offensive | **52.75** | |
| Identity Groups | LLaMA-2 | Toxic | 50.50 | |
| Identity Groups | LLaMA-2 | Abusive | 41.51 | |
| Euphemistic Abuse | LLaMA-2 | Hateful | 34.56 | |
| Euphemistic Abuse | LLaMA-2 | Insulting | 35.41 | |
| Euphemistic Abuse | LLaMA-2 | Offensive | **34.91** | |
| Euphemistic Abuse | LLaMA-2 | Toxic | **35.73** | |
| Euphemistic Abuse | LLaMA-2 | Abusive | 33.78 | |
| Comparisons | LLaMA-2 | Hateful | 39.85 | |
| Comparisons | LLaMA-2 | Insulting | **43.52** | |
| Comparisons | LLaMA-2 | Offensive | **40.22** | |
| Comparisons | LLaMA-2 | Toxic | 38.74 | |
| Comparisons | LLaMA-2 | Abusive | 35.62 | |
| Identity Groups | GPT-3.5 | Hateful | 62.63 | 62.47 |
| Identity Groups | GPT-3.5 | Insulting | 62.09 | 57.71 |
| Identity Groups | GPT-3.5 | Offensive | **64.91** | 55.70 |
| Identity Groups | GPT-3.5 | Toxic | **65.97** | 62.87 |
| Identity Groups | GPT-3.5 | Abusive | 55.65 | 54.58 |
| Comparisons | GPT-3.5 | Hateful | 59.39 | 66.71 |
| Comparisons | GPT-3.5 | Insulting | 64.71 | 65.70 |
| Comparisons | GPT-3.5 | Offensive | **58.90** | 64.95 |
| Comparisons | GPT-3.5 | Toxic | 60.49 | 67.16 |
| Comparisons | GPT-3.5 | Abusive | 57.12 | **68.96** |
| Euphemistic Abuse | GPT-3.5 | Hateful | 60.18 | 57.12 |
| Euphemistic Abuse | GPT-3.5 | Insulting | **60.27** | 57.34 |
| Euphemistic Abuse | GPT-3.5 | Offensive | **59.76** | **55.37** |
| Euphemistic Abuse | GPT-3.5 | Toxic | 57.45 | 58.86 |
| Euphemistic Abuse | GPT-3.5 | Abusive | 60.23 | 57.39 |
| Identity Groups | GPT-4 | Hateful | **84.58** | 84.05 |
| Identity Groups | GPT-4 | Insulting | 83.23 | 83.92 |
| Identity Groups | GPT-4 | Offensive | 81.26 | **82.59** |
| Identity Groups | GPT-4 | Toxic | 82.74 | 83.31 |
| Identity Groups | GPT-4 | Abusive | 79.53 | 83.34 |
| Comparisons | GPT-4 | Hateful | 69.09 | 71.07 |
| Comparisons | GPT-4 | Insulting | 71.74 | 69.30 |
| Comparisons | GPT-4 | Offensive | **75.93** | 70.63 |
| Comparisons | GPT-4 | Toxic | 74.76 | 71.24 |
| Comparisons | GPT-4 | Abusive | 71.55 | 72.31 |
| Euphemistic Abuse | GPT-4 | Hateful | 70.33 | 76.77 |
| Euphemistic Abuse | GPT-4 | Insulting | 76.57 | 75.47 |
| Euphemistic Abuse | GPT-4 | Offensive | **76.16** | **76.38** |
| Euphemistic Abuse | GPT-4 | Toxic | 75.37 | 77.30 |
| Euphemistic Abuse | GPT-4 | Abusive | 72.59 | **79.28** |