ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

TRADUZIONE, INTERPRETAZIONE E INTERCULTURALITA'

Ciclo 37

Settore Concorsuale: 10/L1 - LINGUE, LETTERATURE E CULTURE INGLESE E ANGLO - AMERICANA

Settore Scientifico Disciplinare: INF/01 INFORMATICA

DETECTION AND COMPUTATIONAL ANALYSIS OF INTERNET HATE SPEECH

Presentata da: Aikaterini Korre

Coordinatore Dottorato

Chiara Elefante

Supervisore

Luis Alberto Barrón-Cedeño

Co-supervisore

Beatrice Spallaccia

Esame finale anno 2025

Borsa di dottorato del Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 (CCI 2014IT16M2OP005), risorse FSE REACT-EU, Azione IV.4 “Dottorati e contratti di ricerca su tematiche dell’innovazione” e Azione IV.5 “Dottorati su tematiche Green.” CUP code: J35F21003550006

Acknowledgments

These three years as a PhD student have taught me that, regardless of the topic you choose, the luck you have—good or bad—or even the amount of effort you put, the most important factor is the people around you. Therefore, I would like to take this opportunity to thank those who have not only contributed in this thesis but also played a role in shaping me both as a researcher and as a human being. First and foremost, I would like to thank my supervisor, Alberto Barrón Cedeño, for his constant feedback and guidance throughout the PhD. Secondly, I would like to thank Arianna Muti, Federico Ruggeri, Eleonora Mancini, Eleonora Misino, Francesco Antici, and Andrea Galassi for their collaboration, our many brainstorming sessions, and the reciprocal mental coaching that kept us going. I'm also especially grateful to John Pavlopoulos, who first ignited my interest in Computational Linguistics and is still there always with a useful piece of advice at hand or a new idea. A huge thank you to everyone who annotated any kind of material for me. Believe it or not, without your contribution, this thesis would probably be about half a page long. I cannot help but thank my family—despite not fully understanding what I do and my habit of dodging their questions, they continue to support me unconditionally. Penultimate thanks (I swear!) to all my friends in Greece, Italy, and around the world for helping me take my mind off studying, even just for a bit. Finally, I'd like to thank Lorenzo. Even though he's a buzzer-beater in my thank-you notes, his energy (whether he realizes it or not) has a way of making you feel like everything will turn out alright in the end.

To show you that my procrastination is usually for a good cause, I've prepared another form of appreciation as a thank you. You can find it here, and who knows? Maybe you'll find yourself in there.

Abstract

Hate speech has long been a prevalent issue offline, but with the rise of the internet and social media, its spread has accelerated. The anonymity afforded by these platforms enables individuals to engage in hate speech often without facing substantial consequences. This not only jeopardizes the civility of online communities but also takes a toll on the mental well-being of those targeted. As technology continues to evolve, new opportunities arise to tackle this problem, particularly through the use of natural language processing (NLP). NLP technology can help automate processes that have traditionally been done manually, such as flagging hate speech in online content. Yet, many issues remain unresolved before we can achieve efficient hate speech detection systems. These include foundational challenges, such as defining hate speech, as well as issues that arise at the end of an NLP pipeline, such as evaluating whether a model can generalize when using different data. Additionally, the issue of bias in models poses a significant challenge, as biased training data can lead to inaccurate or unfair results. In this thesis, I will focus on addressing these issues. First, I examine the definitions of hate speech and related concepts like toxicity and abusive language, and their impact on re-annotated datasets. I then compare the original and re-annotated labels in terms of robustness and generalization using a BERT-based classifier. Next, I explore the use of hate speech legislation from three countries for annotation, expanding the task of hate speech detection to prosecutable hate speech detection. The results show even law interpretation can be subjective, which has a consequent effect on model training and evaluation. To address this issue, I introduce a semantic componential analysis of hate speech definitions, leading to the creation of the HateDefCon corpus—450 definitions and an annotation framework for cross-domain and cross-cultural analysis. Staying on the topic of cross-culturality, I present a pipeline for generating parallel multilingual hate speech corpora and discuss the associated challenges. The thesis concludes with an examination of textual biases based on the psycholinguistic aspects of harmful language.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

***Warning**: This thesis contains language that some readers may find offensive or distressing.*

Hate speech is a phenomenon that manifests both offline and online. It is commonly defined as abusive language directed at an individual or group based on specific characteristics, often inciting violence or hostility (Papcunová et al., 2023). While the internet was initially envisioned as a platform to foster dialogue and promote civil discourse (Rheingold, 2000), the freedom of expression it enables has unfortunately contributed to the proliferation of hate speech, undermining those original ideals (Tucker et al., 2018; Mathew et al., 2019).

The rapid technological spread of hate speech calls for equally advanced technological solutions to combat the issue. Technology based on Natural Language Processing (NLP) emerges as a promising approach, surpassing the limitations of simple word filters, which are inadequate for addressing the complexity of the problem (Schmidt and Wiegand, 2017). NLP enables automated methods for identifying and flagging hateful content online, significantly reducing the manual workload of human moderators—a critical advantage given the inherent scalability challenges of manual moderation (Gongane et al., 2022).

However, automated solutions cannot be considered a panacea, as they come with theoretical and practical obstacles. The most fundamental issue is the absence of a universal definition of hate speech. A universal definition of hate speech still stands as an elusive concept (Fino, 2020; Davidson et al., 2017b), as it is inextricably linked with cultural and personal biases (Sap et al., 2022). Additionally, terms related to harmful language, such as toxicity, abuse, offensive language, and cyberbullying, often overlap or exist as hypernyms or hyponyms of one another, complicating the development of precise detection systems (Pachinger et al., 2023). This inconsistency in terminology directly impacts the creation of datasets, as human annotators must label data points (or instances) as harmful or not, based on varying definitions (Fortuna et al., 2020). When different guidelines assign conflicting definitions to the same concept or use the same definition for distinct concepts, the resulting data becomes inconsistent. Such datasets are limited to their specific tasks, and the outcomes they produce are often not reproducible.

This thesis principally addresses the issue of the definitions of hate speech and related concepts such as toxicity and abusive language. It investigates the impact of using varying definitions in different settings, highlighting their effects on both annotation quality, and model performance and evaluation. In addition, this thesis explores a cross-cultural approach to generating hate speech detection datasets. Finally, it addresses issues of bias, triggered by textual tendencies rooted in the data.

1.1 Motivation

The need for more robust methods of hate speech detection, which are both linguistically and culturally aware, has led to the emergence of numerous shared tasks in recent years (Basile et al., 2019; Wiedemann et al., 2020; Wiegand et al., 2019; Struß et al., 2019; Fersini et al., 2022; Pavlopoulos et al., 2022; Mandla et al., 2021). These tasks have encouraged researchers to develop and evaluate their own systems. Despite the significant advancements in classification approaches driven by these shared tasks, critical issues such as definition, bias, and reproducibility remain inadequately addressed.

Regarding annotation, various paradigms have been proposed, each with its own strengths and weaknesses, in an effort to balance objectivity and subjectivity in labeling. Although the trend is moving toward a perscriptivist approach, where all subjective opinions on what constitutes hate speech are considered valid (Cabitza et al., 2023; Rottger et al., 2022; Ruggeri et al., 2023), researchers often arbitrarily select a definition specific to their task without considering the broader implications. This practice can compromise the reproducibility of their work. Moreover, the definitions themselves are part of the problem as they differ significantly depending on the domain or culture they are derived from. These variations can be seen in the length and comprehensiveness of the definitions, as well as their content—some definitions specify target groups, while others do not. Even if a specific and comprehensive definition were established, the issue of bias would still obstruct efficient model development. As I will explore in the following chapters, human bias and model bias are two key factors in an NLP pipeline (including for hate speech detection), with model bias being a direct consequence of human bias.

1.2 Objectives

Having introduced some critical gaps in the research of hate speech detection, the objectives of the thesis are outlined as follows:

1. To investigate how using different definitions of hate speech and related concepts, such as abusive language, offensive language, and toxicity, impact the annotation process, particularly when compared to how existing datasets were originally annotated, and how these variations, in turn, affect model evaluation.

2. To examine hate speech laws as an alternative to conventional hate speech definitions, focusing on the task of prosecutable hate speech detection. This approach is explored because laws are often more thoughtfully considered by experts, and therefore allows us to assess whether annotating with legal frameworks reduces subjectivity and improves inter-annotator agreement.

3. To provide a resource and an annotation framework capable of accommodating as many hate speech definitions as possible, regardless of domain or culture, and facilitating a comparison that enables the selection of the most suitable definition for a given task.

4. To provide an efficient method for generating parallel multilingual hate speech data while preserving toxicity levels and examining how cultural aspects of hate speech are maintained.

5. To examine whether the textual aspects of hate speech, such as inferred emotions or communication style, influence the annotators during the annotation process.

1.3 Contributions

With respect to the objectives that were set above, the following contributions are presented here:

C_1: Re-annotation experiments are conducted using different definitions in each annotation round for the same dataset. The definitions used include hate speech, toxicity, abusive language, offensive language, along with a round where no definition was provided. These experiments demonstrate how varying definitions significantly impact both the annotation process and model evaluation. The results show that higher agreement and better model performance are achieved when using more general definitions (such as toxicity) or when no definition is considered. This highlights the importance of definition choice in hate speech detection and presents re-annotation as a valuable method for re-examining existing datasets to explore how different definitions affect labeling consistency and model outcomes.

C_2: Given that existing definitions of hate speech fall short in providing a robust framework, this research shifts focus to hate speech laws as an alternative. Legal experts are consulted to annotate a hate speech dataset for its prosecutability, and various approaches are explored, including pretrained models trained on both hate speech and legal data, as well as the use of two large language models (Qwen2-7B-Instruct and Meta-Llama-3-70B). To address the time-consuming nature of data acquisition for prosecutable hate speech, pseudo-labeling is employed to enhance the performance of the pretrained models. This study highlights the importance of expanding research on prosecutable hate speech and offers insights into effective strategies for addressing hate speech within legal frameworks. The findings indicate that using definitions does not necessarily provide a more objective alternative to annotating with definitions, while also highlighting that greater attention must be paid to the differences between laws to improve classification.

C_3: A Semantic Componential Analysis framework is proposed for cross-cultural and cross-domain analysis of hate speech definitions. The first dataset of definitions, derived from five domains—online dictionaries, research papers, Wikipedia articles, legislation, and online platforms—is created and analyzed into semantic components. The analysis reveals significant variation in the components across definitions, with many domains borrowing definitions from one another without accounting for cultural context. Zero-shot model experiments are conducted using the proposed dataset, employing three popular open-source large language models to examine the impact of different definitions on hate speech detection. The findings show that large language models are sensitive to the definitions used, with responses for hate speech detection changing according to the complexity of the definitions in the prompts.

C_4: A pipeline is designed to explore the possibility of creating a parallel multilingual hate speech dataset using machine translation. This study evaluates the feasibility of this approach by assessing the quality of the translations, calculating the toxicity levels of both the original and target texts. Additionally, a qualitative analysis is performed to gain further semantic and grammatical insights.

C_5: Artificial intelligence models are applied to two harmful language datasets, Jigsaw's Special Rater Pool and Measuring Hate Speech, to generate probabilities for various text aspects, including inferring the demographic information (age and gender) of the author behind the potentially harmful text, as well as the expressed emotions, emotionality, sentiment, and communication style. A statistical regression analysis is then performed to examine how these text aspects correlate with hate speech and toxicity annotations. The findings confirm that most of the text aspects are correlated with how hate speech and toxicity are perceived by annotators. The study demonstrates that psycholinguistic text aspects, which can be derived from the author's personality, are statistically associated with annotators' perceptions of harmful language and may influence how annotators label the texts.

1.4 Structure

The remainder of this thesis is structured into 8 chapters, with an overview of each chapter provided below:

Chapter 2 provides an overview of the fundamental concepts necessary for a rigorous understanding of this thesis. The concepts discussed include the definitions of hate speech and other related concepts, such as toxic and abusive language, as well as legal approaches to hate speech. The chapter also approaches hate speech through a psycholinguistic lens. Finally, the chapter includes a section on computational methods used in the experimental parts of the thesis, such as pretrained language models and large language models, as well as an overview of the employed metrics for evaluation.

Chapter 3 reviews seminal work on hate speech detection, with a focus on annotation methods and automated approaches. Within these sections, the chapter also highlights the challenges inherent in hate speech detection—challenges that also form the core themes of this thesis: the definitions of hate speech, generalization, and bias.

Chapter 4 introduces re-annotation as a method to assess the robustness and generalization capacity of existing harmful language datasets. The re-annotation process is conducted in multiple rounds, employing crowd-sourced workers and alternating the definitions of harmful language across rounds. The experimental section concludes with an evaluation of the datasets, comparing their performance in a binary harmful language classification task using both the original labels and the re-annotated labels.

Chapter 5 explores the feasibility of using hate speech legislation, rather than conventional definitions, for annotation and model fine-tuning. The chapter begins by introducing the task of prosecutable hate speech detection, followed by a methodology that leverages three country-specific laws (from Greece, Italy, and the UK) to conduct expert annotation of an existing hate speech dataset. It then presents experiments employing pretrained and large language models to detect prosecutable hate speech.

Chapter 6 introduces a novel annotation framework grounded in the linguistic concept of semantic componential analysis. Instead of annotating potential instances, this approach focuses on annotating hate speech definitions using a highly fine-grained methodology. This process reveals cross-cultural and cross-domain variations in the definitions, which are further analyzed by employing the annotated definitions for classification tasks using large language models.

Chapter 7 presents a pipeline for creating parallel multilingual corpora tailored for hate speech detection. The methodology employs machine translation to generate corpora in multiple languages, followed by the use of a pretrained model to produce toxicity scores. These scores are used to assess the preservation of sentence toxicity and to filter the corpus accordingly. The chapter concludes with a qualitative evaluation to validate the results.

Chapter 8 shifts focus to the text itself by exploring various psycholinguistic aspects of harmful language and their impact on annotation. These aspects are represented by predicted values from AI models, and regression analysis is employed to investigate the correlation between these aspects and the labels in the datasets used.

Chapter 9 concludes the thesis with a summary of its key findings, an overarching assessment of the outcomes, and directions for future work.

1.5 Publications

This thesis has resulted in some published work or work that is currently under review. I list the publications or preprints as follows:

- Katerina Korre, John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, Ion Androutsopoulos, Lucas Dixon, and Alberto Barrón-cedeño. 2023. Harmful Language Datasets: An Assessment of Robustness. In The 7th Workshop on Online Abuse and Harms (WOAH), pages 221–230, Toronto, Canada. Association for Computational Linguistics. (*research product of Chapter 4*)
- Katerina Korre, Arianna Muti, and Alberto Barrón-Cedeño. 2024. The Challenges of Creating a Parallel Multilingual Hate Speech Corpus: An Exploration. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 15842–15853, Torino, Italia. ELRA and ICCL. (*research product of Chapter 7*)
- Katerina Korre, John Pavlopoulos, Paolo Gajo, and Alberto Barrón-Cedeño. 2024. Hate Speech According to the Law: An Analysis for Effective Detection. arXiv preprint arXiv:2412.06144. Available at: https://arxiv.org/abs/2412.06144. (*research product of Chapter 5*)
- Katerina Korre, Arianna Muti, Federico Ruggeri, and Alberto Barrón-Cedeño. 2025. Untangling Hate Speech Definitions: A Semantic Componential Analysis Across Cultures and Domains. Findings of the Association for Computational Linguistics: NAACL 2025. (*research product of Chapter 6*)
- Katerina Korre, Seren Yenikent, Angelo Basile, Beatrice Spallaccia, Marc Franco-Salvador, and Alberto Barrón-Cedeño. 2025. Psycholinguistic effects on the annotation of harmful language. Accepted in Language Resources and Evaluation Journal. (*research product of Chapter 8*)

Chapter 2

Background

This chapter outlines the key concepts and terminology essential for understanding this thesis. Section 2.1 provides an overview of the definitions of hate speech, as well as other related terms such as toxicity, abusive language, and offensive language, and clarifying their distinctions, while also discussing the legal and other linguistic dimensions of hate speech. Section 2.2 introduces foundational concepts of supervised learning, including annotation, modeling, and evaluation methods.

2.1 Defining Hate Speech

Identifying hate speech is a challenge due to its multifaceted nature as a linguistic, social and legal topic (Jahan and Oussalah, 2023). This section offers an in-depth overview of terms related to hate speech, including toxicity, abusive language, and offensive language. It further explores current legal frameworks surrounding hate speech and concludes with an examination of its textual characteristics.

2.1.1 Hate Speech and Related Concepts

Definitions of hate speech are often found across four primary domains: (1) legal, (2) lexical, (3) scientific, and (4) practical, each of which can vary in scope and focus (Papcunová et al., 2023). Legal definitions are generally more straightforward than other types, as they aim to identify behaviors that violate existing laws and warrant government regulation. These definitions clarify actions that involve incitement, promotion, or justification of hatred, discrimination, or hostility toward specific individuals or groups based on attributes such as race, ethnicity, religion, disability, gender, age, sexual orientation, or gender identity (Alkiviadou et al., 2020; EU Commission, 2016; UN General Assembly; Council of Europe, 1997). There is still no universally agreed-upon legal definition of hate speech, and each country, interprets hate speech through its own distinct legal framework (Brown, 2017).

This diversity is also evident in the lexical domain, where dictionary definitions of hate speech vary in how many details they provide. Some dictionaries, such as *The Cambridge Dictionary*, offer a detailed definition, describing hate speech as "public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation".[1] Others, like *Merriam-Webster*, provide a broader definition, describing hate speech simply as "speech expressing hatred of a particular group of people", without specifying the characteristics of the group.[2] Despite these differences, all definitions focus on one core idea: hate speech is described as a written or spoken message that conveys hatred and incites violence against a particular group, highlighting the meaning of the collocation between the two words: *hate* and *speech* (Papcunová et al., 2023).

The lack of consensus, as far as the term hate speech is concerned, is evident also in research, in different disciplines, such as economics, philosophy, sociology, psychology, or computer science (Papcunová et al., 2023). As I will show later in this thesis and in the context of NLP, researchers either define hate speech based on their own criteria to suit their research objectives or adopt existing definitions from sources such as dictionaries or existing academic studies.

Defining hate speech becomes even more complex when attempting to distinguish between explicit and implicit forms. Implicit hate speech relies on subtle, indirect linguistic cues that may conceal the aggression or violence. Examples of these linguistic cues include circumlocution, metaphors, and sarcasm, which render this type of hate speech a significant challenge for automated systems (Ocampo et al., 2023; Gao et al., 2017). Although it may appear less direct or hurtful compared to overt abuse, implicit hate speech can be just as damaging as more explicit forms of aggression (Gao et al., 2017).

Practical definitions can be considered primarily those found in the community guidelines of online platforms (e.g., social media such as X, previously known as Twitter)[3] and technological companies (e.g., Microsoft).[4] Many of these companies have signed a Code of Conduct with the European Commission to regulate illegal

[1] https://dictionary.cambridge.org/dictionary/english/hate-speech
[2] https://www.merriam-webster.com/dictionary/hate%20speech
[3] https://x.com
[4] https://www.microsoft.com

hate speech (EU Commission, 2016). For instance, YouTube defines hate speech as "content promoting violence or hatred against individuals or groups based on any of the following attributes: age, caste, disability, ethnicity, gender identity and expression, nationality, race, immigration status, religion, sex/gender, sexual orientation, victims of a major violent event and their kin, and veteran status".[5]

Fortuna and Nunes (2018) conducted a review of all four types of hate speech definitions (legal, lexical, scientific, and practical). After performing a definition and dataset analysis, they proposed their own definition as follows:

Hate speech is language that attacks or diminishes, incites violence or hatred against groups based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity, or other factors. It can manifest in various linguistic styles, including subtle forms or even humor.

The definition created by Fortuna and Nunes (2018) stands out as one of the most comprehensive definitions of hate speech. It covers both the motives and potential consequences of hate speech, provides examples of possible targeted attributes, and takes into account the possibility of implicit expressions of hate.

The discussion about definitions does not revolve only around hate speech, but also other related terms, such as toxic or abusive language. As Parekh (2012) states, it is necessary to distinguish hate speech from related terms that do not fit the definition, such as expressing dislike, lack of respect, a demeaning view of others, disapproval, the use of abusive or insulting speech, and speech that does "not call for action". Yet, and especially in the context of NLP, related concepts like toxicity, abusiveness, offensiveness, incivility, and cyberbullying are often studied in parallel with hate speech, though each might have distinct characteristics.

Toxic language is typically defined as speech "somewhat likely to make a user leave a discussion or give up on sharing their perspective", and can be further broken down into concepts like, disrespect, identity attacks, insults, obscenity, rudeness, threats, unreasonableness (Pavlopoulos et al., 2021). Abusive language can be described as "ascribing a social identity to a person that is judged negatively by a (perceived) majority of society. This identity is seen as a shameful, unworthy, morally objectionable or marginal identity. The target of judgment is seen as a representative of a group and it is ascribed negative qualities that are taken to be universal" (Struß et al., 2019). On the other hand, offensive language can be seen as "any form of non-acceptable language, or a targeted offense, veiled or

[5] https://www.youtube.com/howyoutubeworks/our-commitments/standing-up-to-hate/

direct. This consists of insult/threat to an individual or a group or profanity and swearing" (Zampieri et al., 2019, 2020). Although these definitions may have distinct characteristics, some of them can overlap significantly, and others may be considered hypernyms or hyponyms of one another, leading to terminological ambiguity.

To clarify these concepts and facilitate more precise identification of online hate speech, researchers have attempted to disambiguate these terms. Pachinger et al. (2023) found that in many studies, toxic language serves as an umbrella term encompassing abusive language. Furthermore, instances classified as uncivil often intersect with those identified as toxic or abusive. Incivility, however, is a term more commonly used in social science research. Similarly, Fortuna et al. (2020), in a comparative analysis of datasets and definitions, noted that terms like toxicity, abusiveness, and offensiveness are frequently used interchangeably, further complicating the landscape of harmful language detection. The suggested guidelines from the output of the research conducted on the definitions so far warns researchers to avoid creating redundant definitions, unless extremely necessary, as well as to clearly define their task purposes before selecting a definition (Pachinger et al., 2023; Papcunová et al., 2023; Khurana et al., 2022; Fortuna et al., 2020).

2.1.2 Legal Frameworks of Hate Speech

Forming legal frameworks for hate speech is a challenging task as legislation differs from one country to another. According to Marwick and Miller (2014) there are three main legal scholarly approaches to the definition of hate speech: (i) *content-based*, which includes words, expressions, symbols and iconographies generally considered offensive to a particular group of people and objectively offensive to society; (ii) *intent-based*, which requires the speaker's communicative intention to incite hatred or violence against a particular minority, member of a minority, or person associated with a minority without communicating any legitimate message; and (iii) *harms-based*, that causes the victim harm, such as loss of self-esteem, physical and mental stress, social and economic subordination and effective exclusion from mainstream society. These definitions share at least one core element with one another, which hints the potential for a unified framework.

A lot of steps have been made in the European Union (EU) in the past, such as the *No Hate Speech Movement* by the Council Of Europe,[6] a campaign that mobilises young people to report hate speech and cyberbullying to the relevant

[6] https://www.coe.int/en/web/no-hate-campaign/no-hate-speech-movement

authorities and on social media channels, and the *EU Hate Speech Code*, which required social media platforms to review most notifications potentially flagged for hate speech within 24 hours.[7] After long negotiations, the EU has managed to set up a common regime that requires that all EU Member States make incitement to hatred or violence a punishable act. This is called *The Council Framework Decision* and it has been into force since 2024.[8] The EU Code of Conduct and the EU's new Digital Services Act require also platforms, such as X and Facebook, to take action with regard to illegal content,[9] while in 2018, the introduction of the Audiovisual Media Directive called upon member states to guarantee that audiovisual media services, provided by media service providers and video sharing platform providers under their jurisdiction, to not include any instigation of violence or hostility directed at any group or individual based on the criteria (such as gender) listed in Article 21 of the EU Charter of Fundamental Rights (Flick, 2020).[10]

2.1.3 Linguistic Attributes of Hate Speech

Although hate speech is not a new phenomenon, its proliferation has accelerated in recent years, largely due to the widespread use of the internet.[11] Ironically, while the internet enables the rapid spread of hate speech, it also provides a means for recording and preserving online discourse, which can later be analyzed on a linguistic level (McCulloch, 2019).

Unlike real-time, in-person hate speech, online hate speech displays unique characteristics partly as a result of platform moderation and, generally, the nature of digital interactions. For instance, Retta (2023) shows that explicit slurs are rare in hate speech found in social media, largely due to moderation efforts. More specifically, people tend to use non-slur insults that carry similarly offensive, derogatory, and threatening connotations, in order to evade detection by hate

[7] https://www.theguardian.com/technology/2016/may/31/\facebook-youtube-twitter-microsoft-eu-hate-speech-code

[8] Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law. In the remainder of this paper, we shall refer to this as 'EU law' or 'EU Framework Decision' for simplification.

[9] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).

[10] Directive (EU) 2018/1808 of the European Parliament and of the Council concerning the provision of audiovisual media services (Audiovisual Media Services Directive).

[11] https://www.coe.int/en/web/combating-hate-speech/what-is-hate-speech-and-why-is-it-a-problem-

speech algorithms, while slurs appear more frequently in one-on-one hate speech than in broader discriminatory discourse. Retta (2023) also suggests that in both cases (either using slurs or indirect insults), a divisive “us vs them” mentality is reinforced, which constructs an “us” group as positive and safe, while hinting that “them” is threatening or dangerous.

From a pragmatics point of view, Määttä (2023) shows that there is a tendency in online discourse to hate-normalizing language, contributing to the spread of hateful ideologies as common sense, asserting also that many hate comments carry performative power by reinforcing negative stereotypes and legitimizing hate, highlighting the blurred line between stating and doing harm through language.

Discourse-analysis studies have also focused on the role of emotions in hate speech. According to Alcántara-Plá (2024), emotions in hate speech messages resonate with recipients by framing a threat or opportunity that is personally relevant to them, fostering a sense of potential loss or gain, rather than simply inciting hatred for its own sake. Alcántara-Plá (2024) further distinguishes between “shallow” and “deep objects” in hate speech. “Shallow” objects refer to the immediate, visible targets or victims—such as Muslims, Jews, or the LGBTI community—while “deep objects” involve the underlying fears, beliefs, and historical narratives that shape the discourse. The relationship between these shallow and deep objects is crucial in shaping the emotional impact of hate speech, as counter-narratives are currently just focusing on the former.

2.2 Supervised Learning

In NLP, supervised learning is one of the most widely adopted techniques (Tiwari, 2022). The term “supervised” indicates that a model is trained on a labeled dataset, meaning that each training example is paired with the correct output. The ultimate goal of a supervised model is to learn the mapping from inputs to outputs that can be used to predict on unseen data.

This thesis is concerned with two types of problems that can be tackled with supervised learning:

- **Classification**: Where the output variable is a category, such as “hate speech” or “not hate speech”.

- **Regression**: Where the output variable is continuous, such as predicting the intensity of abusive language, for example, by using a Likert scale.

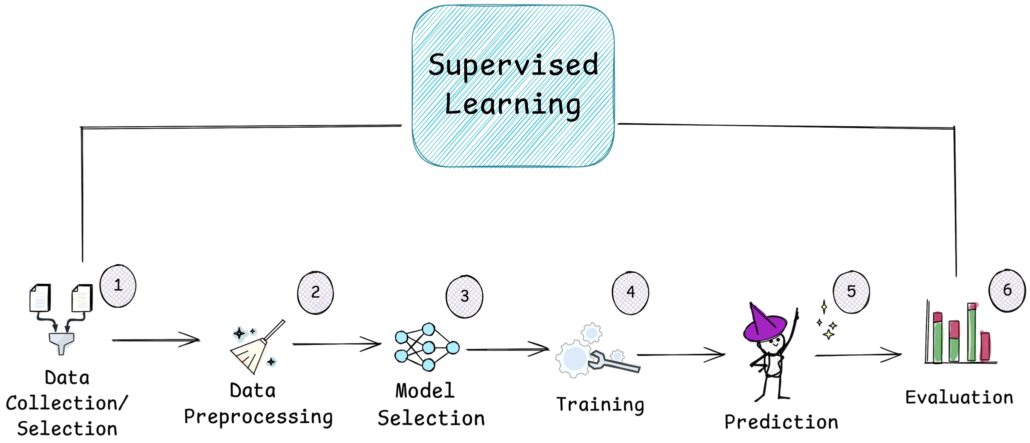


Figure 2.1: Steps for carrying out supervised learning.

The process of training and evaluating a supervised learning model typically follows the steps depicted in Figure 2.1. The first step is *data collection* which involves collecting raw data and labeling them so that the new dataset contains both inputs (features) and outputs (labels) for a given task. In this thesis, a distinction is considered between *data collection* and *data selection* with the latter referring to the selection of a dataset with already labeled data. The second step is to *preprocess* the data. Examples of data preprocessing include removing noise, such as special characters, tokenizing the sentences into words, and lemmatizing. The third step is *model selection*, where the proper algorithm has to be chosen. Older supervised learning algorithms include logistic and linear regression, decision trees, random forest, Naive Bayes, support vector machines (SVM), and, more recently, neural networks (Jiang et al., 2020). The next step is the *training* of the model, which uses the training data to teach the model the relationship between the input and output. Then, the model provides the *predictions* on new, unseen data (test set). The final step is the *evaluation* of the model, in which the model's performance is assessed, using evaluation metrics such as accuracy, precision, recall, F_1 score, mean squared error (MSE) or mean absolute error (MAE).

2.2.1 Annotation

Data annotation is the process of labeling or tagging data with meaningful information that allows a machine learning model to learn from it. In supervised learning, this step is crucial as models rely on annotated datasets to make accurate predictions. Without proper annotation, a model cannot recognize patterns or learn the relationships between input features and the target output.

The quality and accuracy of annotations directly impact the performance of machine learning models. In this section, we explore the importance of annotation, common types of annotation, methods for annotating data, and challenges that come with the process.

There are several methods to annotate data depending on the type of data and the task at hand. The most common methods include:

Expert Annotation This type of annotation requires expert human annotators who label the data. This is typically the most accurate method but can be time-consuming and costly, especially for large datasets (Chau et al., 2020).

Crowdsourcing Crowdsourcing is a method where a large number of workers (often via platforms like Amazon Mechanical Turk)[12] are hired to annotate data. This method is cost-effective and scalable, but the quality of annotations may vary (Gillick and Liu, 2010; Waseem, 2016).

Automated Annotation Automated annotation involves using machine learning models to automatically label data. While faster than manual annotation, it may not be as accurate. However, there is currently a trend towards mixed approaches, which involve using a round of automated annotation and then manual inspection by humans, which can reduce the workload, as well as the initial costs (Smit et al., 2020; Ostyakova et al., 2023).

When it comes to how texts are annotated, practically, Poletto et al. (2019) compare the three main annotation types:

Binary Annotation This type of annotation involves assigning one of two labels to each instance, making it straightforward for manual annotation and computational processing. However, it can oversimplify complex or subjective phenomena, which are common in human language.

Rating Scales These scales, such as the Likert scale, use a range of values to capture varying degrees of a concept, making them better suited for handling subjective opinions compared to binary systems. However, they can still lead to high inter-annotator disagreement, inconsistencies and scale bias.

Best-Worst Scaling This type of annotation involves presenting annotators with a set of items and asking them to choose the best and worst among them based on a specific property. It offers a comparative approach that helps address some issues found in binary and rating scale methods.

[12] https://www.mturk.com/

Ensuring that annotations are consistent and accurate is a major concern, as it refers back to the problem of data quality, bias, and fairness. This problem becomes more challenging with subjective tasks, like hate speech detection or sentiment analysis, where annotators often rely on personal experiences, leading to ambiguity in the data that can complicate assigning a single label. To address this, current research has developed various annotation paradigms that try to mitigate these issues, depending on whether researchers prefer a subjective or objective methodology. The *descriptive* paradigm promotes annotator subjectivity to capture a diverse range of beliefs in datasets, whereas the *prescriptive* paradigm limits subjectivity, enforcing a consistent perspective through strict annotation guidelines, useful for model training consistency (Rottger et al., 2022). The *perspectivist* approach extends descriptivism, opposing aggregated gold standards to preserve individual opinions and cultural diversity, a shift that has inspired new frameworks capturing human subjectivity (Cabitza et al., 2023). Another one is the contrastive model approach to disagreement which integrates multiple hate-speech-related tasks (like identifying aggressive or abusive language) under a single framework as it assumes that these behaviors often share a common foundation (Rizzi et al., 2024). There are also more fine-grained and task-specific frameworks like the one by Kumar et al. (2024) on harmful language and who consider the intersectional and complex nature of triggers that can lead to violence, such as the socio-political context and the speaker's intent.

Metrics for Inter-Annotator Agreement To evaluate inter-annotator agreement, metrics like Cohen's Kappa and Krippendorff's Alpha are widely used. These metrics account for chance agreement, offering a more robust measure compared to simple percentage agreement.

1. **Cohen's Kappa** (κ) quantifies the level of agreement between two annotators while accounting for the agreement that might occur by chance.

   Let:

   - P_o: Observed agreement (proportion of instances where the annotators agree).
   - P_e: Expected agreement by chance (proportion of agreement expected based on the annotators' individual label distributions).

   $$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where:

$$P_e = \sum_k \left(P_1(k) \cdot P_2(k) \right),$$

and $P_1(k)$ and $P_2(k)$ are the proportions of times each annotator assigned the label k.

A κ value of 1 indicates perfect agreement, 0 indicates agreement equivalent to chance, and negative values indicate less than chance agreement.

2. **Krippendorff's Alpha** (α) is a generalized agreement metric that works for any number of annotators, labels, or data types (e.g., nominal, ordinal, interval).

   Let:

   - D_o: Observed disagreement (sum of squared differences between annotations).
   - D_e: Expected disagreement by chance (sum of squared differences expected under random labeling).

   $$\alpha = 1 - \frac{D_o}{D_e}$$

   where:

   $$D_o = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \delta(x_i, x_j),$$

   and D_e is computed similarly but assumes random labeling based on the frequency distribution of labels.

   The function $\delta(x_i, x_j)$ represents the distance between two annotations x_i and x_j. For nominal data:

   $$\delta(x_i, x_j) = \begin{cases} 0, & \text{if } x_i = x_j, \\ 1, & \text{if } x_i \neq x_j. \end{cases}$$

2.2.2 Pretrained Language Models

In recent years, pretrained language models (PLMs) have revolutionized the field of NLP. Classical ML-based NLP models required task-specific features and were often trained from scratch for each application. PLMs, on the other hand, are trained on vast amounts of text data in an unsupervised fashion and can be fine-tuned for specific tasks, saving time and computational resources.

PLMs – with the most important one, BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) –, have achieved state-of-the-art performance in a variety of NLP tasks, such as text classification, named entity recognition, question answering, and more (Sun et al., 2022). BERT's bidirectional approach to understanding context within a sentence has made it a game-changer for the NLP community. Historically, language models relied on simpler approaches, such as n-grams or word embeddings like Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), which learned distributed representations of words in a vector space. While effective, they do not capture meaning in different contexts. BERT addresses this challenge by using a Transformer architecture with a bidirectional approach, which helps it understand context in both directions and learn deep contextual relationships between words. BERT uses a mechanism called self-attention (Vaswani et al., 2017), which allows the model to weigh the importance of different words in a sentence, irrespective of their position and is particularly useful in capturing long-range dependencies and relationships.

BERT is pretrained on a large corpus of text and can then be fine-tuned for specific tasks. The pretraining involves two main objectives:

- Masked Language Modeling (MLM): During pretraining, BERT randomly masks some of the words in a sentence and tries to predict the missing words based on their context. This is a bidirectional process, meaning BERT uses both the left and right context to predict the masked word.

- Next Sentence Prediction (NSP): In addition to MLM, BERT is also trained to predict whether one sentence follows another, which helps it understand relationships between sentences.

BERT's architecture inspired a range of specialized models designed for specific tasks, especially in sensitive and complex areas like hate speech detection and sentiment analysis. These models leverage BERT's transfer learning capabilities, fine-tuning the base model on task-specific datasets to achieve high accuracy in their respective applications. One example is HateBERT (Caselli et al., 2021), which was trained on a large corpus of abusive and toxic language from platforms like Reddit, with a focus on content that included hate speech, offensive language, and harassment. HateBERT is one of the PLMs that are used for the experiments in this thesis in Chapter 5.

2.2.3 Large Language Models

Large Language Models (LLMs) have gathered significant attention due to their remarkable ability to understand and generate human language with exceptional accuracy (Kumar et al., 2018b). Notable examples of LLMs include GPT-4 (OpenAI et al., 2024), Mistral 7B (Jiang et al., 2023), and PaLM 2 (Anil et al., 2023). These models leverage advanced deep learning techniques and the Transformer architecture, similar to Pretrained Language Models (PLMs) like BERT. However, a key distinguishing factor lies in their scale. LLMs are trained on massive datasets comprising diverse and extensive text corpora and are built with billions of parameters, far surpassing the size of traditional PLMs. Furthermore, their autoregressive training paradigm, where models predict the next token in a sequence, allows a very coherent and contextually relevant text generation.

In the previous section, it is mentioned that early PLMs require more explicit fine-tuning, including training on top of the pretraining data, with new, labeled data that are task-specific. LLMs ease the fine-tuning procedure with *prompting*. Prompting refers to the practice of crafting specific inputs, or "prompts" to guide the responses generated by the LLMs. The prompt is, essentially, the text or question posed to the LLM, which the model then uses as a basis for generating relevant, coherent, and contextually appropriate outputs. Prompting can often accomplish what fine-tuning does, especially for task-oriented language tasks, thanks to the model's broad generalization capabilities (Reynolds and McDonell, 2021; Brown et al., 2020). There are many types of prompting techniques. Some of the most popular ones include:

- **Zero-shot prompting**: Simply asking the model to perform a task without examples.

- **Few-shot prompting**: Providing a few examples in the prompt to guide the model toward the desired output style.

- **Instruction tuning**: Fine-tuning with a wide array of task instructions to generalize the model's ability to follow new instructions accurately.

Yet, just like PLMs, LLMs can inherit biases from the data they are trained on, perpetuating, and amplifying harmful social biases (Gallegos et al., 2024). In addition, because LLMs are complex black-box models, it is difficult to create protocols to evaluate their performance, as well as detect the existence of potential biases and to understand the reasons behind certain outputs or model reasoning (Chang et al., 2024; Zhao et al., 2024; Huang and Chang, 2023).

2.2.4 Evaluation Metrics

This thesis is concerned with classification, regression, and similarity problems, such as those encountered in tasks like machine translation or text generation, where the similarity between a reference sentence and a generated sentence must be calculated. This section presents the evaluation metrics used for these types of problems.

Metrics for Classification A classification task involves a model classifying an input instance by predicting its class label. In some of the experiments of this thesis, the classification tasks are in a binary setting, for example, whether an instance is "hate speech" (HS) or "not hate speech" (Not HS). Conventionally, the metrics used for the evaluation of classification are: accuracy, precision, recall, and F_1 score. The metrics are explained below using as an example a binary hate speech detection setting, along with their mathematical formulas. To understand the metrics, we first need to define the confusion matrix, which shows the counts of each type of classification outcome. An example of a confusion matrix is found in Table 2.1.

1. **Accuracy** represents the overall proportion of correct predictions. It is preferred when the data classes ("hate speech", "not hate speech") are balanced.

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

2. **Precision** indicates how many of the instances predicted as hate speech were actually hate speech.

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

	Predicted HS	Predicted Not HS
Actual HS	TP	FN
Actual Not HS	FP	TN

Table 2.1: Example of a confusion matrix, where **True Positives (TP)** are the instances of hate speech that the model correctly identifies as hate speech; **False Negatives (FN)** are the instances of hate speech that the model incorrectly identifies as not hate speech; **False Positives (FP)** are the instances of not hate speech that the model incorrectly identifies as hate speech; **True Negatives (TN)** are the instances of not hate speech that the model correctly identifies as not hate speech.

3. **Recall** measures how many of the actual hate speech instances were correctly identified.

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

4. $\mathbf{F_1}$ **score** balances precision and recall, providing a single metric that is useful when there is an uneven class distribution.

$$\text{F}_1 \text{ score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Metrics for Regression To evaluate the performance of NLP models on regression problems, such as predicting the intensity of hate speech or abusive language, we can use Mean Absolute Error (MAE) and Mean Squared Error (MSE).

Let:

- n: Number of data points, for example the total number of social media comments in the dataset.

- y_i: Actual value for the i-th data point. In the case of predicting the intensity of hate speech, it would be the actual observed intensity score for hate speech in the i-th comment (it could be a continuous value, e.g., from 0 to 1, where 0 represents neutral language and 1 represents extremely abusive or hateful language)

- $\hat{y}_i$: Predicted value (i.e., intensity score) for the i-th data point, generated by the model.

1. **Mean Absolute Error (MAE)**: MAE represents the average absolute difference between predicted and actual values. In this case, it tells us the average deviation of our predictions from the actual hate speech intensity levels.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

2. **Mean Squared Error (MSE)**: MSE penalizes larger errors more heavily than MAE, making it more sensitive to outliers. MSE penalizes larger errors more than smaller ones because it squares the error for each prediction. This metric is particularly useful if we want the model to be more sensitive to high-intensity wrong predictions (e.g., strongly abusive comments with high intensity scores).

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Metrics for Similarity To calculate the similarity between an automatically generated or translated sentence and a reference sentence, BLEU score and BERTScore are commonly used.

1. **BLEU Score (Bilingual Evaluation Understudy)** measures the similarity between a generated text and a reference text by comparing their overlapping n-grams (e.g., sequences of 1, 2, 3,..n words). It is a precision-based metric but also accounts for brevity to avoid favoring very short outputs.

   Let:

   - p_n: Precision for n-gram matching (proportion of n-grams in the candidate text that match n-grams in the reference text).
   - BP: Brevity penalty, which penalizes outputs shorter than the reference text.
   - N: Maximum n-gram size to consider (commonly 4).

$$\text{BLEU} = BP \cdot \exp\left(\frac{1}{N}\sum_{n=1}^{N} \log(p_n)\right)$$

where:

$$BP = \begin{cases} 1 & \text{if } c > r, \\ \exp\left(1 - \frac{r}{c}\right) & \text{if } c \leq r, \end{cases}$$

and c and r are the lengths of the candidate and reference texts, respectively.

2. **BERTScore** evaluates the semantic similarity between generated and reference texts by comparing their contextual embeddings, obtained using a pretrained transformer model like BERT. Unlike BLEU, it measures semantic rather than surface-level similarity.

   Let:

   - $E(\text{candidate})$: Set of embeddings for the tokens in the candidate text.
   - $E(\text{reference})$: Set of embeddings for the tokens in the reference text.
   - $\text{sim}(e_i, e_j)$: Cosine similarity between embeddings e_i and e_j.

   Precision (P) and Recall (R) are computed as:

$$\text{Precision (P)} = \frac{1}{|E(\text{candidate})|} \sum_{e_i \in E(\text{candidate})} \max_{e_j \in E(\text{reference})} \text{sim}(e_i, e_j)$$

$$\text{Recall (R)} = \frac{1}{|E(\text{reference})|} \sum_{e_j \in E(\text{reference})} \max_{e_i \in E(\text{candidate})} \text{sim}(e_i, e_j)$$

   Finally, the BERTScore F_1 is computed as:

$$\text{BERTScore} = F_1 = 2 \times \frac{\text{P} \times \text{R}}{\text{P} + \text{R}}$$

Chapter 3

Related Work

This chapter reviews related work on hate speech detection. Specifically, Section 3.1 explores prior research on annotation methodologies for hate speech, focusing on annotator selection, defining the task scope, and potential biases. Section 3.2 examines approaches leveraging pretrained language models (PLMs) and large language models (LLMs), that were introduced in Chapter 2, for hate speech detection purposes. It also explores multilingual strategies and addresses the challenge of generalizability.

3.1 Annotation Methods for Hate Speech

Annotating hate speech comes with distinct challenges, as hate speech is inherently subjective, shaped by personal, cultural, and contextual biases of annotators (Sap et al., 2022). A significant body of work has been dedicated to exploring strategies for annotating hate speech, with questions revolving around annotator selection (e.g., experts or crowd-sourced workers), and the design of annotation schemes defined by the scope of the task. This section synthesizes key contributions in the field, focusing on their methods and the trade-offs they address.

3.1.1 Annotator Selection

Annotator selection is one of the foundational steps in the annotation pipeline. Choosing what type of annotation to follow in a given study is essentially a question of priorities, namely the quality-cost trade-off. There is a wide variety of strategies with regard to the number and the background of the annotators, as described in Chapter 2 and also summarized by Poletto et al. (2021) into three main options: (a) having data annotated by experts, (b) having them annotated by amateur/non-expert annotators recruited either as volunteers (often among students) or on a crowdsourcing platform—like Amazon Mechanical Turk—or, (c) using an automatic classifier to assign labels. In general, expert annotations are preferred and considered of better quality. Waseem (2016), shows that amateur annotators are more likely to label items as hate speech than expert annotators, while systems trained on expert annotations tend to outperform those trained on amateur annotations. Similarly, Lopez Long et al. (2021) highlight a tendency among crowd workers to overuse the abusive class, which creates an unrealistic class balance and negatively impacts classification accuracy. With the emergence of LLMs, more cost-efficient approaches have started to be adopted, by automatically annotating the data. Tan et al. (2024) provide a survey on this new phenomenon, highlighting that this method can work in supervised learning scenarios, for example by employing LLMs as both data annotators and learnable models, and iteratively fine-tune LLMs in their self-annotated data, as done by Huang and Chang (2023).

In this thesis, we employ all aforementioned annotator selection methods. In Chapter 4, we employ crowd workers to annotate harmful language providing them with different definitions, while in Chapter 5, we employ both experts and LLMs to create pseudo-labels aiming to improve model performance for prosecutable hate speech detection.

3.1.2 Defining the Task

A recurring theme across the literature is the difficulty in defining hate speech as a clear-cut category, which, consequently, poses a challenge when defining annotation guidelines. According to Davidson et al. (2017a), no formal definition exists. With that in mind, studies adopt definitions depending on their task formulation. Assimakopoulos et al. (2020) argue that hate speech lies on a continuum of discriminatory discourse, often conveyed indirectly through linguistic subtleties. Recognizing that annotators may differ in their thresholds for identifying such content, they suggest avoiding the direct use of the label hate speech and they propose a multilayer annotation scheme informed by critical discourse analysis.

In contrast, Ron et al. (2023) adopt an approach which deconstructs hate speech into a set of distinct discursive aspects. Unlike Assimakopoulos et al. (2020), who emphasize the continuum of discriminatory discourse, Ron et al. (2023) propose a more modular approach that breaks down hate speech into its components. While their annotation scheme can capture some aspects of hate speech, its focus on antisemitism limits its generalizability to broader contexts. Another modular approach is the one presented in Khurana et al. (2022), who develop hate speech criteria with input from legal and social science perspectives to help researchers create precise definitions and annotation guidelines. They propose five criteria: target groups, dominance, perpetrator characteristics, type of negative group reference, and potential consequences/effects. Rather than prescribing a single definition, they offer a meta-prescriptive, modular approach, allowing for adjustments based on specific tasks.

Other attempts to define hate speech draw inspiration from the legal field. For instance, Zufall et al. (2022) operationalized whether a post falls under criminal law by translating the EU Framework Decision 2008/913/JHA into a sequence of binary decisions. Their results suggest that annotation is more efficient when breaking the decision into two subtasks: target group detection (which groups are referred to in a post, for example minorities, such as women and immigrants) and conduct detection (what punishable action patterns are referred to in a post). For their purposes, they create a Punishable Hate Speech Dataset in German, with a moderate inter-annotator agreement (IAA) for target group and targeting conduct (Cohen's Kappa 0.52 to 0.70), but low for the punishable label (0.33 to 0.43). Another attempt to exploring hate speech laws from an NLP outlook was that of Fišer et al. (2017), which presented the legal framework, a dataset, as well as an annotation schema for unacceptable discourse practices in Slovene, which includes annotations for the typology and target of the unacceptable discourse. More recently, Luo et al. (2023) introduced hate speech detection grounded in enforceable legal definitions. They create a gold standard dataset annotated by legal experts and establish baseline LLMs, incorporating explanations. In Chapter 5, we present our approach to addressing the issue of prosecutable hate speech, touching upon topics such as the subjectivity of annotators and providing a cross-cultural analysis of legal definitions of hate speech.

The issue of the definitions is evident not only with respect to hate speech detection, but also generally in harmful language detection. Recent NLP work has focused on comparing harmful language definitions (abusiveness, toxicity, offensiveness, etc.) rather than comparing hate speech definitions per se. Discussing two important terms 'trolling' and 'flaming', KhosraviNik and Esposito (2018) very eloquently suggest that "[d]espite the widespread (and often overlapping) use of these two terms, the utmost complexity of the discursive practices and behaviours of online hostility has somehow managed to hinder the development of principled definitions and univocal terminology". This inconsistency is present in published datasets of harmful language. For that reason, Fortuna et al. (2020) focus on clarifying applied categories of harmful language and homogenizing different datasets by treating class representations as vectors rather than just relying on their definitions. They additionally evaluate a classifier's (Perpective API (Lees et al., 2022)) ability to distinguish standardized categories from non-harmful messages through binary classification. As their final remark, the authors stress the importance of providing guidelines for annotation schemas and avoiding the creation of new categories unless absolutely necessary, with clear examples and justifications when new categories are introduced. In a consequent attempt to disambiguate harmful language definitions used in NLP research, Pachinger et al. (2023) review and compare definitions of uncivil, offensive, and toxic comments across 23 papers from various fields, aiming to foster more unified scientific resources. Both the works of Fortuna et al. (2020) and Pachinger et al. (2023) highlight the need for consistent terminology to promote clarity in the scientific research of hate speech and harmful language, in general.

In Chapter 6, we offer a resource that includes a wide range of hate speech definitions —from general to specific, and from various sources— allowing researchers to select from established, culturally relevant options. This helps to avoid the creation of custom definitions when unnecessary, which could introduce bias into experimental models, beginning with the annotation process.

3.1.3 Annotator Bias in Harmful Language Detection

Previous research highlights the role of the annotator and their characteristics, i.e., internal and contextual factors belonging to the individuals who rate the comments in terms of hate speech (Sap et al., 2022; Waseem, 2016). Specifically, in dataset creation and annotation, these characteristics can be translated into a high bias level that is specific to the annotators who label each instance. In fact, bias can be observed even within the same group of annotators (such as minority groups), since "members of the marginalized communities may still show bias towards their own communities" (Goyal et al., 2022). Bias in hate speech datasets can have detrimental effects because currently these human-annotated datasets are still the key component for building AI systems. Specifically in the case of abusive language detection, models have the tendency of being biased towards identity words of a certain group of people because of imbalanced training datasets. For example, Park et al. (2018) note that phrases such as "You are a good woman" are marked as ‘sexist’ probably because of the word ‘woman’, underlining that this unfair bias can inhibit the robustness of models and therefore their efficient practical use. In turn, the resulting systems can propagate biased opinions, by straying away from an objective point of view. By integrating social scientific theories, Davani et al. (2023) show that indeed hate speech classifiers trained on human annotations include biases toward historically marginalized groups. To reduce these biases, they suggest diversifying annotation teams and modeling annotation biases rather than treating them as mere noise. With regard to modelling and incorporating sociodemographic information into multi-annotator models for toxicity detection, Orlikowski et al. (2023) show that this method does not significantly improve performance. However, while sociodemographic attributes may not provide additional benefits universally, their careful use could still be valuable depending on the task and model architecture.

Al Kuwatly et al. (2020), on the other hand, explored annotator bias using classifiers trained on data from demographically distinct annotator groups (gender, education, native language, age group) exploiting corpora from Wikipedia’s Detox project (Wulczyn et al., 2017). Their results show that, although there can be bias in the corresponding group annotations, it also depends on the definitions of hate speech used for the annotation and experiments, underlining the need for a deep examination and understanding a priori. Other studies have focused on resolving the bias issue by studying a limited number of demographic groups. For example, Sap et al. (2022), concentrated on different identity groups as well as groups that differ with regard to their beliefs. More specifically, they first conducted a *breadth-of-workers* controlled study, where they collected ratings of toxicity for a set of 15 manually curated posts from 641 annotators of different races, attitudes, and political beliefs. They also proceeded to a *breadth-of-posts* study, replicating a standard toxic language annotation by collecting toxicity ratings for approximately 600 posts, from a diverse pool of 173 annotators. They discover that the perceptions of toxicity are indeed affected by annotators’ demographic identities and beliefs and that there are several associations when isolating specific text characteristics. Goyal et al. (2022) form groups of raters (i.e. LGBT, African American, and a control group, which comprised people that did not identify with either of the two former groups) and they assign a toxicity annotation task. Specifically, they explore how raters’ self-described identity impacts how they annotate toxicity in online comments. They calculate some descriptive statistics, as well as conducting a regression analysis, where they find significant differences among the groups. They suggest using special rater groups under certain circumstances. The study of Goyal et al. (2022) is the basis of Chapter 8 in which regression analysis is used to examine the correlation between inferred demographics and other text aspects (such as emotionality and sentiment) and the labeling of harmful language. Text aspects are discussed in the next section.

3.1.4 Text Aspects and Annotator Perception

Text aspects are usually perceived by the individual, while many computational linguistics studies have focused on how easily these aspects are perceptible in order to create ground-truth datasets for NLP purposes. For example, demographic information about the author, such as gender, can also be perceived by the raters during annotation, however, always running the risk of bias. Nguyen et al. (2014) perform in-language and cross-language annotation experiments by asking raters to guess the gender of the authors of tweets. The in-language experiments yielded accuracy of 70.5%, while in the cross-lingual experiments, the accuracy drops at 60%, indicating that cultural identity influences how raters perceive certain demographic aspects of the text. On the other hand, emotions can also be perceived, but it also depends on the type of emotion. More specifically, Strapparava and Mihalcea (2010) create an emotion dataset using Ekman emotions and they calculate IAA by averaging the annotations in a pairwise manner and using Pearson's correlation. The results showed that the annotators agreed more on the emotions of fear and sadness (∼68% and ∼63% , respectively), followed by joy (∼59%), anger (∼49%), and disgust (∼44%). On the annotation for sentiment analysis, Bobicev and Sokolova (2017) highlight the difficulty of the task by presenting the IAA scores from previous work, which do not exceed ∼75%, while also conducting further experiments in multiclass, multi-label sentiment annotation of messages, achieving an average agreement of 79%. Communication styles can also be perceived by the annotators. In Štajner (2021), the IAA for those communication styles was calculated as the percentage of instances with perfect agreement with Emotionality. The perfect agreement for *fact-oriented* is 52%, for *self-revealing* 63%, for *action-seeking* 73%, and for *information-seeking* 80%. Finally, they also calculate emotionality with 53% perfect agreement.

All of the aforementioned studies show that text aspects are perceivable during annotation. However, there is always a degree of subjectivity, which derives from personal background and life experiences, that influences the labeling process (Pei and Jurgens, 2023). Chapter 8 of this thesis delves into different text aspects (inferred gender, age, emotionality, sentiment, and communication style of the author) to show if they correlate with how annotators label harmful texts.

3.2 Hate Speech Detection

This section explores various approaches to hate speech detection, focusing particularly on PLMs—primarily BERT-based models—,as well as LLMs. Additionally, it examines relevant research on the topic of generalizability and provides an overview of cross-lingual approaches to hate speech detection.

3.2.1 PLM Approaches for Hate Speech Detection

Transformers models (see Chapter 2) have proven to be the most effective approach for hate speech detection, consistently surpassing other methods in various studies (Ramos et al., 2024). This increase in transformers approaches overlaps with the second OffensEval task (Zampieri et al., 2020), a shared task on offensive language identification,[1] where most participants started shifting from older deep learning approaches (such as LSTMs) to transformer models, rendering them the state-of-the-art at that time (Ramos et al., 2024).

The flexibility and adaptability of BERT-based models, enhanced by the creation of specialized variants and hybrid techniques, have greatly contributed to hate speech detection efforts. Several studies have fine-tuned BERT for hate speech detection. For example, Arcila-Calderón et al. (2022) fine-tune BERT to detect hate speech in Greek, Spanish, and Italian, outperforming other deep learning and machine learning models. Similarly, Saleh et al. (2023), show that BERT-based classifiers perform better compared to other deep learning approaches, due to the large amount of data BERT is pretrained. Examining the context of offensive language, Casavantes et al. (2023) employ BERT contextual vectors, among other techniques to consider text and metadata showing that they provide a clear advantage regarding a traditional approach of only using text. Jahan et al. (2021) show that BERT-based ensemble methods, i.e., combining multiple models to enhance performance by aggregating their predictions, yield optimal results.

Furthermore, re-training BERT for specific tasks is another common approach, including for the task of hate speech detection. An example is HateBERT, a BERT model re-trained to detect abusive language in English. It was trained on RAL-E, a large-scale dataset of English Reddit comments sourced from communities banned for offensive, abusive, or hateful content (Caselli et al., 2021). Moreover, there is BERT-HateXplain, another BERT-based model fine-tuned on annotated data from

[1] https://sites.google.com/site/offensevalsharedtask/home

HateXplain, a benchmark dataset for hate speech detection (Mathew et al., 2021). This dataset includes rationales for identifying hate speech and offensive language, comprising 20,000 posts from Gab and Twitter. BERT-HateXplain demonstrates improved performance and better bias handling compared to standard BERT models in benchmarking. Similarly, RoBERTa-based models, a more robust version of BERT (Liu et al., 2019), have been employed for hate speech detection (Alonso et al., 2020). Vidgen et al. (2021a) used a human-and-model-in-the-loop process to train an online hate detection system, leveraging RoBERTa and dynamically collected expert-annotated datasets with fine-grained labels and perturbations to improve detection performance.

Finally, language-specific and multilingual models have proven effective in non-English contexts. For example, AlBERTo (Polignano et al., 2019a), a model tailored for Italian, has been fine-tuned for hate speech detection in this language (Polignano et al., 2019b). Jahan et al. (2022) use FinBERT, the Finnish version of BERT (Virtanen et al., 2019), for hate speech detection in the Finnish language. Another popular approach is employing mBERT (Devlin et al., 2019), a multilingual BERT model, that has been fine-tuned for cross-linguistic tasks, and allows the detection of offensive content across multiple languages when fine-tuned, as in the case of Bigoulaeva et al. (2023).

3.2.2 LLMs for Hate Speech Detection

LLMs have marked a new age for classification and detection tasks. The capabilities of LLMs have turned research to focus more on explainability, which applies also to the task of hate speech detection. Roy et al. (2023) use various prompt variations and input information to evaluate LLMs in a zero-shot setting, without adding in-context examples. Three LLMs (GPT-3.5, text-davinci, and Flan-T5) and three datasets (HateXplain (Mathew et al., 2021), implicit hate (ElSherief et al., 2021), and ToxicSpans(Pavlopoulos et al., 2022)) are selected for evaluation. They find that including target information in the pipeline improves model performance by approximately 20-30% over the baseline across datasets. On a similar note, Plaza-del arco et al. (2023) explore zero-shot learning with prompting for hate speech detection, investigating its effectiveness in detecting hate speech across three languages with limited labeled data. Their findings emphasize the importance of prompt selection on performance and suggest that prompting, especially with recent LLMs, can match or even surpass fine-tuned models, making it a promising alternative for under-resourced languages.Yang et al. (2023) introduce a hate speech detection framework, *HARE*, which leverages the reasoning capabilities of LLMs to explain hate speech to help users understand its harmful effects, enhancing the supervision of detection models. Experiments on the Social Bias Inference Corpus (Sap et al., 2020) and Implicit Hate benchmark (ElSherief et al., 2021) show that their method, which uses model-generated data, consistently outperforms baselines that rely on existing human annotations. Their analysis also reveals that this approach improves explanation quality and enhances generalization to unseen datasets.

With respect to LLMs for hate speech detection, there is a current turn towards in-context learning, i.e., where an LLM directly performs a new task without any update to its parameters by taking a test instance, new task description and a few training examples (e.g. input-label pairs) as its input. Han and Tang (2022) adopt this method and use GPT-3, exploring ways to create effective prompts to enhance performance for hate speech detection. They show that a significant amount of input-label pairs is essential for achieving strong performance, and that detailed task descriptions can further improve outcomes by incorporating our prior knowledge to guide inference. Jin et al. (2024) introduce GPT-HateCheck, a suite of functional tests for hate speech detection generated and validated by large language models. Their empirical results showed that the generated examples are more diverse and natural than the templated-based counterparts introduced in previous work.

3.2.3 Generalizability in Hate Speech Detection

One of the most frequently discussed problems is the inability of models to generalize, namely the fact that models underperform when tested on a test set from different source than the training set (Swamy et al., 2019; Karan and Šnajder, 2018; Gröndahl et al., 2018). This is a problem which also harmful language detection systems have to face and which aggravates when the models have to deal with more than one language or in the case of implicitness, i.e., “abusive language that is not conveyed by (unambiguously) abusive words” (Wiegand et al., 2021). With regard to the former case, Yin and Zubiaga (2021) claim that, when models are applied cross-lingually, there is a performance drop that indicates that model performance had been severely over-estimated as testing on the same dataset the training set derived from is not a realistic representation of the distribution of unseen data. Attempts to improve the performance models involve merging seen and unseen datasets, using transfer learning, (i.e., leveraging pre-existing knowledge of the model to improve performance) and re-labeling (Talat et al., 2018; Karan and Šnajder, 2018). However, in the majority of cases, instances from the source dataset are needed to achieve high performance (Fortuna et al., 2021). In addition, various characteristics of datasets have been examined as variables for an effective generalization, including the work of Swamy et al. (2019), who suggested that more balanced datasets are healthier for generalization, and that datasets need to be as representative as possible of all facets of harmful language, in order for detection models to generalize better.

When considering implicit forms of abusive language (including hate speech), Wiegand et al. (2021) highlight that very few resources exist, and that very little is known about the implicit content in existing abusive language datasets. For that reason, they create a taxonomy of subtypes of implicit abusive language that enables a more fine-grained approach to detecting such cases. Some of these subtypes (e.g., dehumanization and stereotyping) are addressed in Chapter 6 as components of hate speech definitions.

Since there are variations of harmful language, topic generalization is also seen as a gateway. Chiril et al. (2022) digressing from standard binary approaches, they perform cross-dataset knowledge transfer with different topical focuses and targets (e.g., racism, sexism, xenophobia), showing that training on multiple topic-specific datasets helps models generalize better than using a single, topic-generic dataset. They also show that using a multitask approach, where models learn to detect both hatefulness and the specific topic, enables finer-grained and more adaptable hate speech detection, making it easier to generalize to new, unseen data, as well as that injecting domain-independent affective resources (like SenticNet, EmoSenticNet, and HurtLex) helps models generalize better to specific hate speech expressions, especially in cases where annotated data for new topics or targets is missing. Bourgeade et al. (2023) propose a set of topic oriented analyses of the generalizability of harmful language datasets (on misogyny, sexism, racism, xenophobia, etc.), and they show how topic-generic and topic-specific datasets yield different degrees and nature of generalization, finally suggesting that the use of mixtures allows smoothing out individual weaknesses. Ludwig et al. (2022) analyzed how well hate speech classifiers generalize to different target groups under controlled conditions, focusing on the HateXplain dataset. They found that naive classifiers exhibit bias toward specific target groups, but unsupervised domain adaptation can improve cross-target generalization, though results depend heavily on the amount and type of training data. Other methods of aiding in generalization includes incorporating sentiment or emotion analysis like in the case of Hong and Gauch (2023) who found that emotion knowledge enhanced generalizability, particularly for pretrained models not adapted to the detection of harmful language.

The challenge of generalizability is discussed in Chapter 4, where we experiment with different definitions of harmful language, in order to examine the ability of a BERT-based classifier, as well as the robustness of harmful language datasets.

3.2.4 Cross-lingual Approaches to Hate Speech Detection

Recently, there has been an emergence of cross-lingual approaches in hate speech detection. Given the limited resources in several languages, one-shot and few-shot are two of the most preferred approaches (Mozafari et al., 2022; Zia et al., 2022; Tita and Zubiaga, 2021; Stappen et al., 2020). Another preferred approach is knowledge transfer that can be enhanced with augmentation methods. For example, Pamungkas and Patti (2019) implemented a hybrid approach with deep learning and a multilingual lexicon to cross-domain and cross-lingual detection of abusive content. Bigoulaeva et al. (2022) used cross-lingual word embeddings to train neural network systems on a source language and apply it to a target language to make up for the lack of labeled examples. They also incorporate unlabeled target language data for further model improvements by bootstrapping labels using an ensemble of different model architectures. Arango et al. (2021) propose a hate specific data representation (i.e., hate speech word embeddings) and evaluate its effectiveness against general-purpose universal representations most of which, unlike their proposed model, have been trained on massive amounts of data. They focus on a cross-lingual setting, in which one needs to classify hate speech in one language without having access to any labeled data for that language. Bigoulaeva et al. (2021) used bilingual word embeddings-based classifiers and they achieve good performance on the target language by training only on the source dataset. Using their transferred system, they bootstrap on unlabeled target language data, improving the performance of standard cross-lingual transfer approaches. They use English as a high-resource language and German as the target language for which only a small amount of annotated corpora are available. Their results indicate that cross-lingual transfer learning together with their approach to leverage additional unlabeled data is an effective way of achieving good performance on low-resource target languages without the need for any target-language annotations.

When considering cross-cultural approaches, it is essential to acknowledge and account for the impact of cultural differences on annotation and translation. Lee et al. (2023a) delve into how individuals from different countries perceive hate speech, introducing CReHate, a cross-cultural re-annotation of the sampled SBIC dataset (Sap et al., 2020). This dataset includes annotations from five distinct countries: Australia, Singapore, South Africa, the United Kingdom, and the United States. Their statistical analysis highlights significant differences based on nationality, with only 59.4% of the samples achieving consensus among all countries. In a separate study, Lee et al. (2023b) attempt to quantify the cultural insensitivity of three monolingual (Arabic, English and Korean) hate speech classifiers by evaluating their performance on translated datasets from the other two languages, showing that hate speech classifiers evaluated on datasets from other cultures yield significantly lower F_1 scores, up to almost 50%. Compared to their study, Chapter 7 of this thesis focuses on the initial machine translation step. More on cross-cultural awareness, Hershcovich et al. (2022) highlight the necessity of addressing the lack of cross-culturality in NLP and explore existing strategies to pave the way for a solution. They pinpoint three key areas for mitigating cross-cultural disparities: data collection, model training, and translation. They emphasize the importance of diverse annotation, understanding the trade-off between generalization and adaptation in model usage, and the limitations of reference-based evaluation methods, advocating for culture-sensitive human evaluation. The approach of the study presented in Chapter 7 is based on the three areas outlined by Hershcovich et al. (2022). However, the methodology diverges as the main aim is to automate certain aspects of the parallel data generation pipeline by minimizing reliance on human annotators through the use of translation.

3.3 Conclusion

This chapter provided an overview of the literature on hate speech detection, focusing on key aspects of the task. First, annotation methods are discussed, highlighting that annotator selection and the definitions employed remain contentious topics. These issues are explored in detail in Chapters 4 and 5, where I experiment with re-annotation using both hate speech definitions and related legislative frameworks. The aim is to evaluate which alternatives yield better results in terms of IAA, model evaluation, and generalization. A major contribution of this thesis is the definition resource `HateDefCon`, presented in Chapter 6, alongside an annotation framework designed for cross-domain and cross-cultural definition comparison, touching upon the issue of both the definitions and cross-culturality. Regarding cross-cultural approaches, this thesis also introduces a parallel hate speech dataset generation pipeline in Chapter 7, addressing the challenges inherent to such an approach.

Lastly, while issues of bias are addressed throughout the thesis, Chapter 8 focuses specifically on the textual characteristics of harmful content that may trigger bias and influence annotators' perceptions, as discussed in the present chapter.

Chapter 4

Harmful Language Datasets

An Assessment of Robustness

Many forms of harmful language impact social media despite efforts —legal and technological— to suppress it.[1] A robust solution has yet to emerge, as detecting online harmful language is called to face many challenges, including the absence of a universal definition—necessary for a generalizable approach—or persistent biases, both of which are discussed in Chapter 2.

From blanket terms, such as abusiveness and offensiveness to sub-categories, such as misogyny and cyber-bullying, researchers have explored many variations of harmful language. However, this begs the question of how to select and compare the possible definitions, especially when some categories allow for more accurate predictions for cross-dataset training than others (Fortuna et al., 2021). This chapter focuses on how using different definitions during annotation affects the robustness of harmful language datasets, and their consequent utilization for fine-tuning. A re-annotation of existing datasets with a range of definitions is performed and is later replicated to assess robustness. A qualitative error analysis on the re-annotations follows, which shows that even instances that contain potentially harmful terms might not be perceived as harmful by annotators, underlining the subjectivity of the task. Finally, the generalizability of the existing datasets across the different definitions is analyzed by training BERT-based classifiers and testing them on the original annotations and the re-annotations, concluding that evaluating on broader definitions can yield higher accuracy.

At this point, it must be noted that the term *harmful language* is used as a

[1] https://edition.cnn.com/2022/06/14/asia/japan-cyberbullying-law-intl-hnk-scli/index.html

wildcard term that can be potentially replaced with any term that refers to harmful language, such as toxicity and abusiveness. Section 4.1 presents the re-annotation strategy. In Section 4.2 the experimental setup for training and evaluating with the original and the re-annotated datasets is presented. Finally, the chapter concludes with a discussion of the results and an assessment of the contribution in Section 4.3, concluding with an overall summary in Section 4.4

4.1 Methodology

This section discusses the methodology of this study which is divided in two parts. The first part investigates whether closely-related definitions have an effect on inter-annotator agreement while the second part examines the compatibility and versatility of the present datasets by using them to train models.

4.1.1 Annotation Experiments

In order to study the effect of the definition on inter-annotator agreement, a re-annotation of harmful language datasets is conducted by using alternating definitions and by repeating the annotation in rounds for robustness.

Datasets This study is inspired by the work of Fortuna et al. (2020) on the empirical analysis of hate speech datasets (see Sections 2.1 and 3.1). The initial idea for this study was to use the same data used in Fortuna et al. (2020) in order to produce comparable results. However, not all of the datasets could be used, as the classes used would make it harder for the models to generalize since they refer to specific target groups. For example, the AMI dataset (Fersini et al., 2018) refers specifically to women and the HatEval dataset (Basile et al., 2019) to immigrant minorities. Therefore, the final selection of datasets includes **Davidson** (2017a), **TRAC-1** (Kumar et al., 2018b), and **Toxkaggle** (Jigsaw, 2019). It must also be noted that, for this research, the Davidson dataset is split into two subsets: **DavidsonHS** (for hate speech) and **DavidsonOFF** (for offensiveness), as the two classes correspond to two different definitions of harmful language.

Data Compilation For the purposes of the annotations, 5 different batches of data that contain instances from all aforementioned datasets are created. Each batch contains an equal number of different instances from each dataset, while the

Dataset	Annotation Procedure	Classes	Source
DavidsonHS (Davidson et al., 2017a)	Begining with the hatebase lexicon then CrowdFlower, users coded each tweet (minimum number of annotations per tweet is 3 , sometimes more users coded a tweet when judgments were determined to be unreliable by CF).	Hate speech (25), Not-Hate Speech (25)	Twitter
DavidsonOFF (Davidson et al., 2017a)	Same as above	Offensiveness (25), Not-offensiveness (25)	Twitter
TRAC-1 (Zampieri et al., 2019; Kumar et al., 2018b,a)	The annotation was done using the Crowdflower platform but by what is known as 'internal' annotators in the Crowdflower lingo. The whole of annotation was done by 4 annotators – all of them were native speakers of Hindi, with a nativelike competence in English and were pursuing a doctoral degree in Linguistics.	Overtly Aggressive (OAG) (13), Covertly Aggressive (CAG) (12), Non-Aggressive (NAG) (25)	Facebook
Toxkaggle (Jigsaw, 2019)	Not provided.	Threat (3), Identity hate (3), Severe Toxic (3), Insult (3), Obscene (4), Toxic (9), NonToxic (25)	Wikipedia

Table 4.1: Basic description of dataset. This table was inspired by a similar table found in Fortuna et al. (2020). Davidson et al. (2017a) dataset was split into two separate datasets as Hate Speech and Offensiveness are considered different definitions.

instances are also shuffled. The total number of instances of each of the batches was 200 (out of which we randlomly selected 80 as test questions, for quality control).[2] In each batch we keep a balanced distribution between positive and

[2] Test questions are sample questions that are used to pilot an annotation experiment and assess the quality of the annotators in crowd-sourcing scenarios (Barthet et al., 2023).

negative instances (i.e., harmful language vs non harmful language), while we also keep the balance among the classes derived from each dataset, following the suggestions of Swamy et al. (2019) for better generalisation. Information about class distribution is presented in Table 4.1.

Term	Definitions of harmful language	Citation
TOXIC	A rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion.	(Jigsaw, 2019)
ABUSIVE	Hurtful language, including hate speech, derogatory language and also profanity	(Founta et al., 2018)
OFFENSIVE	Containing “any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct.”	(Zampieri et al., 2019)
HATE	Expressing hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group. In extreme cases this may also be language that threatens or incites violence.	(Davidson et al., 2017a)
HOTA	*Any* of the following: Hateful, Offensive, Toxic, Abusive language (HOTA)	Ours

Table 4.2: The terms and definitions of harmful language that were provided to the annotators during re-annotation.

Annotation Procedure The annotation procedure consists of five annotation experiments, each relating to a different definition for potentially harmful content. For the annotation, we used crowd-sourcing via the Appen platform.[3] The guidelines for the annotations can be found in Appendix A. Since this project was carried out in collaboration with Jigsaw, [4] the raters were compensated according to the company’s regulations, namely a compensation above minimum wage for the annotator region (USA), based on estimates of time to task completion. Jigsaw’s regulations with regard to Appen annotations include reviewing feedback from raters to insure that the task is considered doable and that the raters feel they are compensated fairly.

[3] https://appen.com/

[4] Jigsaw is a unit within Google that explores threats to open societies, and builds technology that inspires scalable solutions.https://jigsaw.google.com/

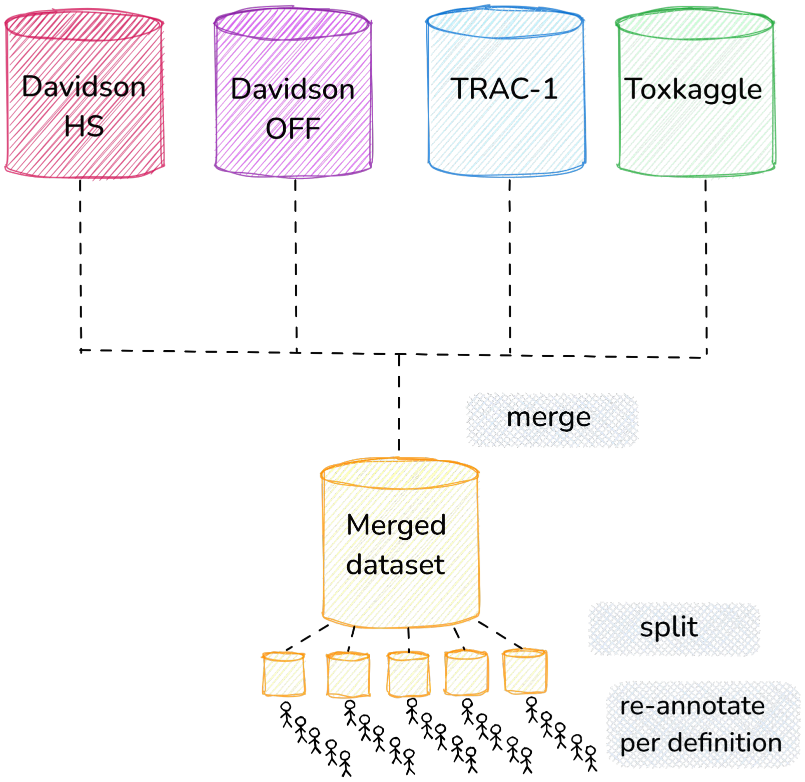


Figure 4.1: Data compilation and annotation procedure. Instances from the 4 datasets were used to create a new dataset that would later be divided into 5 annotation batches.

Each re-annotation experiment was repeated 5 times with different data each time. This variation in the data helps to ensure that the results are not specific to a particular dataset and can be generalized.

Regarding the guidelines, annotators were instructed to read carefully the given definition and examples, and decide whether each text was harmful or not according to the definition provided. The same examples were provided to the annotators across all annotation experiments, and the only thing changed was the term and the definition of harmful language, presented in Table 4.2. Since we used crowd-sourcing, each batch is not necessarily annotated by the same annotators. The quality of the annotators was ensured provided they answer correctly the aforementioned test questions. The data compilation and annotation procedure is also summarized in Figure 4.1.

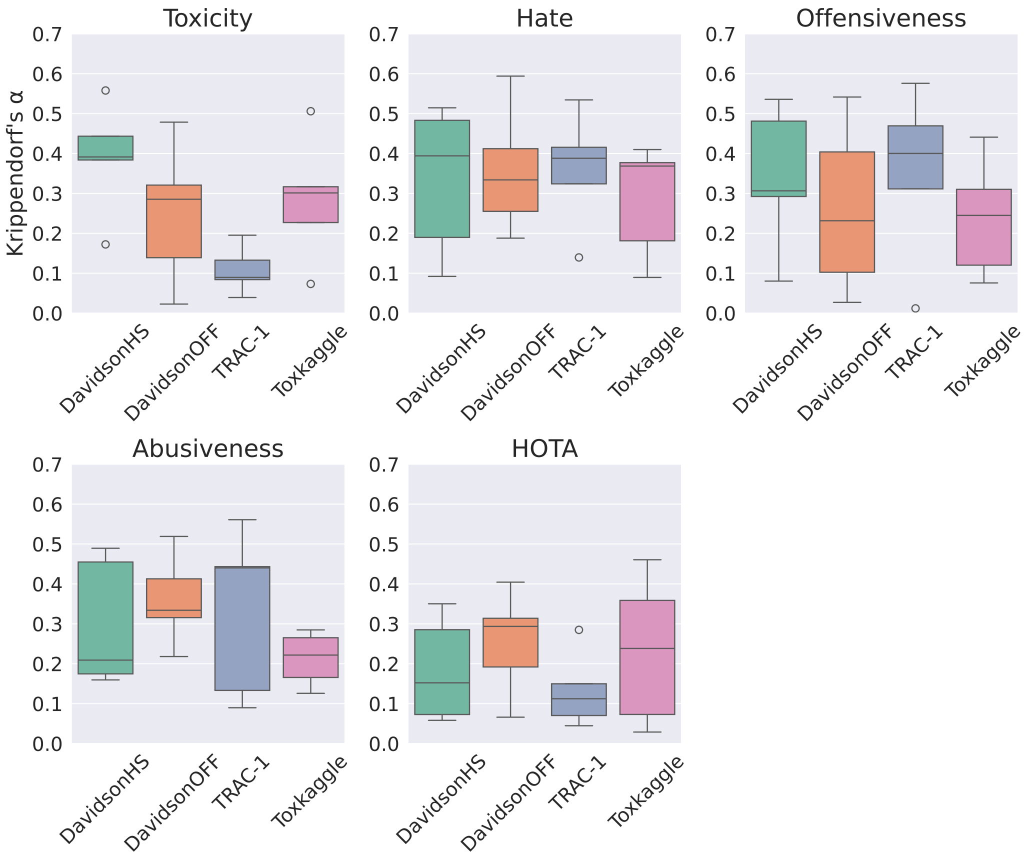


Figure 4.2: Boxplots showing Krippendorf's alpha inter-annotator agreement. The y axis shows the Krippendorf's alpha values while the x axis shows the different datasets. Each plot refers to a different definition.

4.1.2 Annotation analysis

An initial exploratory analysis of the results of the annotation not only shows low inter-annotator agreement in general but also inconsistency both across datasets and across repetitions. This is illustrated in Figure 4.2, which presents boxplots of Krippendorff's alpha inter-annotator agreement, with each boxplot corresponding to a different definition used during the annotation process. Among the 5 definitions, Toxicity and HOTA (see Table 4.2 for the acronym explanation) show more consistent annotation despite the low inter-annotator agreement, which is under 0.5. This poses the question of whether we should trust high inter-annotator agreement and potential inconsistency among repetitions or accept a lower but more robust inter-annotator agreement. Moreover, looking at the inter-annotator agreement per dataset, we see that instances of datasets that were originally annotated with a given definition present a more consistent annotation when re-annotated with another definition. For example, we would expect DavidsonHS to have a more consistent inter-annotator agreement when annotated for hate speech, but we see that it is when it is annotated for toxicity that the result is more robust. Similarly, DavidsonOFF presents slightly more consistent results when annotated for hate speech and abusiveness rather than offensiveness.

	Example	**Source**
1.	*according to you my place is in kitchen. Excuse me while I take out the trash*	Davidson
2.	*It's because you can't bend me, you can't buy me and you can't make me into your nigger!*	Davidson
3.	*California's biggest retards*	Davidson
4.	*"Just because she has light skin doesn't mean anything... I know some Asian with really dark brown skin but that doesn't make them any less Chinese"*	Toxkaggle
5.	*and we shud also destroy taj mahal... qutub minar ....laal qila aftr that visiters seen only GBRoad vd hvng a fun vd ur neighbours*	TRAC-1

Table 4.3: Texts from the evaluation sets with the highest variance of inter-annotator agreement.

Annotation variance can be used to isolate instances with high disagreement. Table 4.3 presents a subset out of the 10 instances with the highest variance per definition that were sampled for the analysis. When annotated for toxicity, these posts included forms of irony. For instance, the example of the 1st row is possibly written by a woman, which might mean that the intention is not to be toxic but to cauterize misogynistic behaviours. In addition, many posts contained vocabulary that is associated with negative sentiments, such as "crazy", "cheater", and "hate". With regard to abusive language, annotators disagreed even for instances that present raw profanity ("bitch", "cocksucker"), potential racism as seen in the 2nd example of the table, and ableism as seen in the third. Similarly, when annotating for offensiveness, the raters did not necessarily annotate positively an instance that contained profanity. Also, racist instances that do not contain obscenities might have been trickier to classify. For example, the author of the 4th example resorts to ostensibly logical reasoning that might disguise the racism that pervades the sentence. Compared to the other definitions that were given during the re-annotation, the sampled re-annotations for hate speech did not show any clear pattern possibly because the definition of hate speech is more restricting referring to specific target groups. However, the same holds true for HOTA, which was the broader term during the re-annotation. The sample that we checked during this qualitative analysis included profanity, references to homosexuality or racism and misogyny, as well as instances that did not contain any harmful language. Noteworthy is also the fact that the sentence in Example 5 appeared with high variance in 3 out of 5 definitions, possibly because of the mixed language use and modified words.

4.2 Experimental setup

The main focus of the experiments is to determine to what degree, and under which conditions (i.e., with which harmful language definition-annotated dataset), the models yield more robust results and generalize better. Two experiments were conducted:

Experiment 1: Testing on the original ground-truth The goal of this experiment is to evaluate the ability of the BERT-based models trained on the original datasets to perform on test sets that rely on the original ground truth annotations. This setup allows to assess how effectively models trained on the original specific annotations can generalize when tested against the labeling schemes of the original datasets.

Experiment 2: Testing on the re-annotations as ground-truth This experiment evaluates the classifiers by testing them on ground truth derived from the re-annotations. The focus here is to explore whether the fine-tuned model aligns more with the re-annotations rather than the originally annotated ground-truth.

4.2.1 Datasets

We use the same four datasets that were used in re-annotation (Davidson et al., 2017a; Kumar et al., 2018b; Jigsaw, 2019) to perform harmful language classification (toxicity/hate speech/offensiveness/abusiveness). More specifically, we first

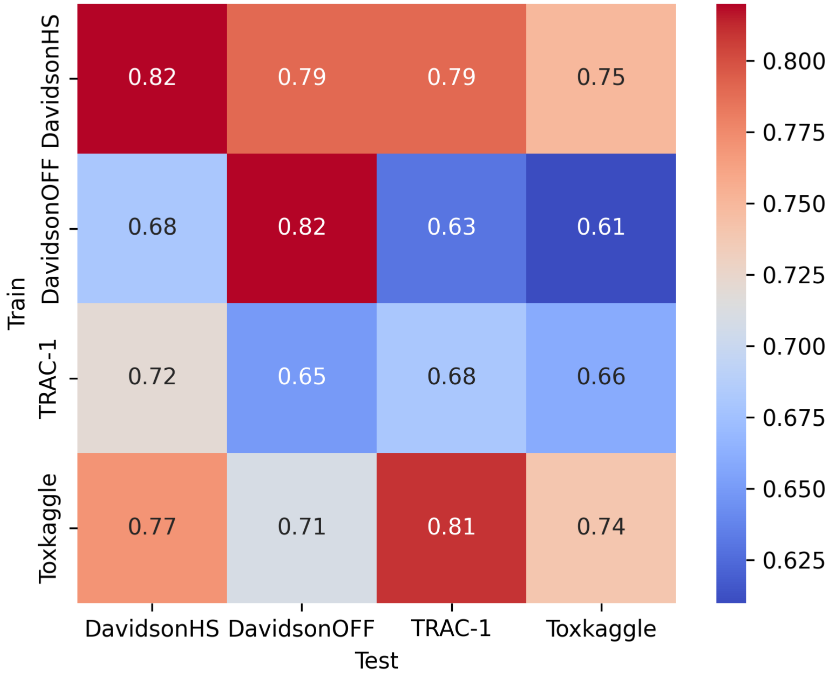


Figure 4.3: Heatmap showing the accuracy on the different test sets using the original ground truth (horizontally) when the model is trained on each corresponding dataset (vertically).

extracted the 1,000 (200 per definition) instances used for the human annotation from the original datasets. Then, with the remaining instances we created 4 balanced datasets that contained an equal amount of positive and negative instances (2650 in total). The evaluation of the model was carried out by calculating the accuracy with respect to the original annotation labels and the ones produced for the new annotation.

4.2.2 Model training

The experiments used the four re-annotated datasets: DavidsonHS, DavidsonOFF, Trac-1, and Toxkaggle. The ktrain library for text classification (Maiya, 2020)[5]

[5] https://github.com/amaiya/ktrain

		Evaluation (re-annotated)			
Training	**Definition**	**DavidsonHS**	**DavidsonOFF**	**TRAC-1**	**Toxkaggle**
DavidsonHS	Toxicity	0.75 (-0.07)	0.69 (-0.10)	0.75 (-0.04)	**0.83** (+0.08)
	Hate Speech	0.64 (-0.18)	0.59 (-0.20)	0.58 (-0.21)	0.59 (-0.16)
	Offensiveness	0.64 (-0.18)	0.64 (-0.15)	0.56 (-0.23)	0.62 (-0.13)
	Abusiveness	0.63 (-0.19)	0.57 (-0.22)	0.62 (-0.17)	0.59 (-0.16)
	HOTA	**0.76** (-0.06)	**0.78** (-0.01)	**0.78** (-0.01)	0.82 (+0.07)
DavidsonOFF	Toxicity	**0.64** (+0.04)	0.72 (-0.10)	0.83 (+0.20)	**0.75** (+0.14)
	Hate Speech	0.58 (-0.10)	0.63 (-0.19)	0.59 (-0.04)	0.59 (-0.02)
	Offensiveness	0.50 (-0.18)	0.66 (-0.16)	0.56 (-0.07)	0.59 (-0.02)
	Abusiveness	0.57 (-0.11)	0.61 (-0.21)	0.62 (-0.01)	0.60 (-0.01)
	HOTA	0.62 (-0.06)	**0.76** (-0.06)	**0.84** (+0.21)	0.74 (+0.13)
TRAC-1	Toxicity	0.67 (-0.05)	0.59 (-0.06)	0.50 (-0.18)	0.53 (-0.13)
	Hate Speech	0.69 (-0.03)	**0.66** (+0.01)	0.53 (-0.15)	0.63 (-0.03)
	Offensiveness	0.69 (-0.03)	0.63 (-0.02)	**0.55** (-0.13)	**0.66** (=)
	Abusiveness	0.70 (-0.02)	0.64 (-0.01)	0.51 (-0.17)	0.65 (+0.01)
	HOTA	**0.71** (-0.01)	**0.66** (+0.01)	0.47 (-0.21)	0.57 (-0.09)
Toxkaggle	Toxicity	0.73 (-0.04)	0.67 (-0.04)	0.77 (-0.04)	**0.85** (+11)
	Hate Speech	0.68 (-0.09)	0.67 (-0.09)	0.63(-0.18)	0.61 (-0.13)
	Offensiveness	0.67(-0.10)	0.69 (-0.02)	0.63(-0.18)	0.68(-0.06)
	Abusiveness	0.71 (-0.06)	0.65 (-0.06)	0.63 (-0.18)	0.61 (-0.13)
	HOTA	**0.79** (+0.02)	**0.77** (+0.06)	**0.81** (=)	0.83 (+0.08)

Table 4.4: Accuracy of BERT trained per dataset (1st column), using the original annotations, and evaluated on our re-annotations per definition. In parentheses is the accuracy increase (green) or decrease (red) compared to the scores obtained on the evaluation data with the original annotations (Figure 4.3).

was employed, using the BERT-base-uncased model for fine-tuning. A batch size of 32 was set, and the default hyperparameters provided by the ktrain library, such as the learning rate, were applied. Fine-tuning was conducted for up to 7 epochs, with early stopping implemented using a patience value of 4. The maximum token length for inputs was set to 100 to align with the characteristics of the datasets.

4.2.3 Results

We assess the classifiers using both the original and the re-annotated ground truth. **Using the source annotations** as our evaluation ground truth, the accuracy of the classifiers is presented in Figure 4.3. We observe that when the model is trained on DavidsonHS datasets, it reaches an accuracy of more than 0.75 in all test sets. As expected, the accuracy is higher when the model is also tested on DavidsonHS.

When the model is trained on DavidsonOFF the accuracy is high only when tested again on DavidsonOFF. Training with Toxkaggle results in more than 0.70 accuracy in all test sets, with the highest accuracy in the TRAC-1 test set (0.81). TRAC-1, on the other hand, shows the lowest accuracy across all test sets (0.65-0.72), with the highest accuracy obtained when testing on DavidsonHS.
Using the re-annotations as the evaluation ground truth, is shown in Table 4.4. The classifiers did not manage to generalize across datasets consistently, which is shown by the fact that accuracy decreases, in comparison to the scores obtained when the original annotations were used for testing our models. There are sparse exceptions where the accuracy increases, for example, when training on Toxkaggle and testing on re-annotations of HOTA, where results were equal (TRAC-1) or better (DavidsonHS, DavidsonOFF, Toxkaggle). In general, the highest accuracy, although still low in terms of what current language models can achieve, is achieved when testing either on the toxicity or HOTA re-annotations. Excluding Toxkaggle, however, we observe that accuracy deteriorated in our re-annotations even when evaluating on test sets derived from the same source as the training set, except for TRAC-1 that presents a slight increase of 0.01 when testing on hate speech and HOTA.

4.3 Discussion

This study offers a new perspective on existing harmful language datasets and urges the NLP community to re-examine current labeling practices through re-annotation. The experiments in this thesis highlight the subjectivity of annotations and emphasize that different types of harmful language cannot be treated as equivalent categories. This complements the work of Pachinger et al. (2023), discussed in Section 3.1, who reveal that many categories used for harmful language either overlap or represent hypernyms or hyponyms of one another.

Taking into account the existing literature (Fortuna et al., 2020; Karan and Šnajder, 2018; Swamy et al., 2019; Yin and Zubiaga, 2021), this study confirms that models face a serious difficulty generalizing, while they can perform better when tested on the same dataset they were trained on, as happens with the two *Davidson* datasets. Yet, our results show some other promising aspects when it comes to model generalization for toxicity detection purposes, as well as building robust datasets through a transparent annotation procedure. More specifically, we see that the models perform better in the two most general definitions, i.e., Toxicity and HOTA. This can be due to pragmatic reasons, namely, classifying items using broad definitions can be an easier task for both the annotators and the trained models. On the other hand, it might be a matter of compatibility between the training data and the testing data. For example, the classes used in the re-annotation procedure were more similar to the ones used in the two *Davidson* sub-sets and *Toxkaggle*, while they were more different compared to *TRAC-1* for which also another definition was originally used (Aggressiveness) that we did not include in our experiments. Focusing on such differences among different datasets could enable researchers to outline the DOs and DON'Ts for annotations and dataset compilation. According to Fortuna et al. (2020), fine-grained toxicity categories are not the optimum option, while more general categories yield better results. Considering that, for the purposes of this experiment we tried to binarize and simplify the datasets, as much as possible, by separating the Davidson dataset and by merging the subcategories in *TRAC-1* and *Toxkaggle*. However, this did not help the performance when it comes to *TRAC-1*. One possible reason behind this could be the fact that *TRAC-1* contains implicit aggressiveness that is harder to detect, even when the model is trained on the respective dataset. The difficulty to detect implicit aggressiveness or other forms of toxicity is not only true for models, but also for human annotators as we saw in Section 4.1.2.

4.4 Conclusion

In spite of recent advances, model generalization in harmful language detection still has significant limitations, particularly when applied across diverse datasets, which have been originally annotated with different definitions. This part of the thesis highlights the critical issues surrounding harmful language definitions, model generalization, and dataset compatibility.

First, a re-annotation experiment was conducted with publicly available datasets, employing crowd-sourced annotators. This experiment revealed the subjectivity inherent in harmful language annotation. Broader definitions, such as Toxicity and HOTA, resulted in more consistent annotations, suggesting that simplicity and generality in annotation guidelines lead to improved robustness.

Second, I trained BERT-based classifiers using the same datasets. Consistent with prior research, the experiments demonstrated that models perform better when tested on the same datasets they were trained on. This was particularly evident in the DavidsonHS and DavidsonOFF datasets, where training and testing compatibility resulted in higher accuracy. However, the models struggled to generalize effectively when tested on datasets with different definitions or implicit forms of harmful language, such as TRAC-1, which focuses on aggressiveness.

The fact that both annotators and models cope better with broader categories should not necessarily be seen as an advantage as that might mean that annotation and models cannot capture more subtle details provided by the definitions. This issue is also discussed in Chapter 6 of this thesis.

These findings emphasize the importance of creating more transparent and pragmatic annotation procedures to build robust datasets. They also highlight the need to establish best practices for dataset compilation, including the use of broader definitions and careful consideration of class compatibility. Re-annotation is a way which can be exploited to do so.

The next chapter will zoom in on a specific category of definitions: definitions of hate speech within national legislation. It will examine whether such definitions—grounded in criteria established and enforced by governmental authorities—can improve the processes of annotation and classification of hate speech detection.

Chapter 5

Hate Speech According to the Law

One of the most important challenges for hate speech detection highlighted in the previous chapters is the universal inability to agree on a single definition of hate speech (Pachinger et al., 2023; Khurana et al., 2022; Fortuna et al., 2020; Davidson et al., 2017b; Fino, 2020), obstructing the creation of generalizable solutions both regarding legislations and NLP applications. The specific issue stems from the subjectivity that lies in the interpretation of hate speech due to different cultural or individual life experiences (Sap et al., 2022; Waseem, 2016). Especially in today's multicultural societies, it is necessary to be aware "of the contexts, and of the power relations involved in balancing hate/free speech, as it is crucial in the analysis of regulation, legislative or other, as well as in any normative debate" (Maussen and Grillo, 2014).

Given that definitions can be subjective, this chapter turns our attention to hate speech laws. Guillén-Nieto (2023) explains that hate speech as a social and legal notion are inherently connected. Socially, hate speech contributes to conflicts and conditions favorable to hate crimes. Hate speech on online platforms often precedes or coordinates racially motivated attacks.[1] For instance, Joshua Bonehill-Paine was prosecuted for racially harassing an ex-Labour MP.[2] Legally, it is seen as "an abstract endangerment statute", balancing the rights to freedom of expression and dignity. Therefore, the assumption is that legislation might provide us with insights and more robust definitions for hate speech detection that are actually culturally informed. The research questions that concern this chapter are:

[1] https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons

[2] https://www.theguardian.com/uk-news/2016/dec/07/racist-troll-guilty-harassing-labour-mp-luciana-berger-joshua-bonehill-paine

1. How does the annotation of hate speech using laws, as compared to conventional definitions (derived from dictionaries or research or customized for a specific task), influence expert inter-annotator agreement, and does this effect vary across countries with different hate speech legislation?

2. Are the discrepancies in inter-annotator agreement reflected in the performance of PLMs and LLMs in hate speech detection?

3. Given the challenging and time-consuming nature of manually creating datasets for prosecutable hate speech, can LLM-generated instances be used to improve the performance of PLMs in detecting prosecutable hate speech?

As in the previous chapter, this one is concerned with both annotation and model experiments. The annotation experiment uses three hate speech legislations from distinct countries, specifically Greece, Italy, and the United Kingdom. For our model experiments, diverse PLMs are employed (Section 5.1.2), while also exploiting zero-shot and few-shot methods with two LLMs: Qwen2-7B-Instruct (Yang et al., 2024) and Meta-Llama-3-70B-2-9b (AI@Meta, 2024) (Section 5.1.3). Subsequently, Qwen2 is also used to automatically generate labels for an additional 1000 instances, to be used to improve the performance of the PLMs (Section 5.2.3). The contributions can be summarized as follows:

- One of the outcomes of this study is the first cross-country, expert-annotated dataset on prosecutable hate speech, considering the **legal ramifications** of published hate speech content and examining its legal compliance.

- Through a **cross-country analysis** of hate speech, it is shown that inter-annotator agreement varies based on the legislation used during annotation. Interestingly, this variation does not necessarily overlap with the model evaluation, suggesting that PLMs and LLMs may carry their own inherent biases.

- **LLM-generated data** do not improve PLM performance, highlighting the still-present necessity of human annotation or a human-in-the-loop approach for prosecutable hate speech detection.

5.1 Method

The focus of this study is on the laws of Greece, Italy, and the UK, for prosecutable hate speech detection purposes. The laws are retrieved from the Global Handbook of Hate Speech Laws, which provides a translated version of the laws into English (see Appendix B.1).[3] The method involves, (1) obtaining expert judgements on a sample of HatEval (Basile et al., 2019) on prosecutable hate speech; (2) exploiting state-of-the-art hate speech PLMs, and LLMs, using zero-shot, few-shot techniques and leave-one-out cross validation; and (3) generating silver labels for an additional 1000 instances for improving the PLMs.

5.1.1 Data and Annotation

For the annotation, instances from HatEval (Basile et al., 2019) are used, a dataset in English which contains hate speech instances both against immigrants and women, annotated via crowd-sourcing. A total of 100 hateful instances were extracted from the dataset.

A re-annotation was conducted with the help of three experts. Expert A is a legal expert, currently a PhD candidate in Criminal law. Expert B is a master's student in Criminology. Expert C holds a master degree in European Law. All of them possess a near-native level of English. They all annotated the sentences voluntarily after examining the three laws (Greek, Italian, and British) by providing labels about whether the given instances were **prosecutable** hate speech or not. There was also a section for comments that allowed the detection of some ambiguous or controversial cases. Each annotation round, which involved evaluating the 100 examples, took around 10 hours to complete for all three laws. The experts reported that they needed 3 hours to read and study the document on the laws and make a summary. They reported that they had to study additional legislative material to be able to judge responsibly. Judging the instances took about 7 hours in total. According to the experts, this is not an easy task. Some tweets were easy to understand and assessing whether the hate speech in them could be subject to prosecution was straightforward. Some others required a significant amount of research, which involved the interpretation of the meaning of the hashtags and abbreviations, as well as the disambiguation of slang and ambiguous words according to the interpreted intent and context of the post. For example, there were cases where the instance could be liable to be sued in court but only for insult or defamation rather than hate speech. This type of annotation is not feasible by other means, such as crowd-sourcing.

Table 5.1 reports the agreement measures. The three experts labeled the instances differently, with Experts B and C being the pair with the highest agreement

[3] https://futurefreespeech.com/global-handbook-on-hate-speech-laws/

National	Cohen's kappa			Class: is prosecutable?			
Law	A & B	B & C	C & A	Yes	Likely	Unlikely	No
Greece	0.20	0.25	0.12	4	16	32	48
Italy	0.31	0.59	0.37	9	16	17	58
UK	0.39	0.61	0.30	8	11	22	59

Table 5.1: Cohen's kappa and the number of instances per ordinal class according to the different country laws.

(0.59 for Italy and 0.61 for the UK). The poor to slight agreement among the pairs of experts suggests inconsistency in their assessments of whether something is prosecutable according to the Greek law, with an average agreement of 0.19. For the Italian and the UK laws, the agreement is moderate, with an average agreement of 0.42 and 0.43, respectively. Hate speech datasets typically present low inter-annotator agreement, due to the subjective nature of the task as well as its dependance on the annotation scheme (Casula and Tonelli, 2020). The results also point to the subjectivity of law interpretation —even by experts— a frequent phenomenon that, among others, often leads to the appealing of court decisions (Donato, 1988). The challenge of achieving a high inter-annotator agreement is also explained in Braun (2023) and in Wacholder et al. (2014) where they emphasize the inherent subjectivity of legal annotation, as each individual interprets the law in a different manner, adding that in many cases the inter-annotator agreement is not even reported.

The judgement variation relates to the main criterion for considering an instance of hate speech liable under the law, which is the fact that the offensive or aggressive verbal act must **incite violence** on the basis of a social or minority characteristic, e.g., gender, religion, race. In that case, and due to the lack of context and other verbal markers, it proved hard for the experts to clearly perceive the possibility of whether those instances could in fact incite violence.

Additionally, the disagreement arises from the subjective global knowledge or experiences of each expert. For example, in the case of the tweet, "Deport them all! #BuildThatWall", two experts felt that the wording and call for deportation did not necessarily reflect racial motives. However, the third expert, taking into account factors such as the timing of the tweet, the political climate, and the implied reference to the US–Mexico border wall, interpreted the tweet as prosecutable hate speech. These differing perspectives do not necessarily imply that one interpretation is wrong. Instead, it highlights the need for clear guidelines in the task—specifically, whether background information, such as political context, should be considered or not. In summary, the disagreement can stem from varying perspectives on issues like the potential for incitement to violence, the distinction between defamation and hate speech, and how one connects the instance to current events.

Taking the above into consideration, this study is aligned with the opinion of Basile (2020) who claims that when it comes to annotating highly subjective tasks such as hate speech, the opinions of all the experts can be correct and that disagreeing annotations that come from diverging opinions should be equally considered in the construction of a gold standard dataset. Due to those differences, and with respect to the opinions of all the experts, the task is formulated as an ordinal regression problem. Cases where all experts label positive are considered **prosecutable hate speech**, cases where two experts agree that an instance is prosecutable are considered **likely prosecutable**, cases where only one expert claims that an instance is prosecutable are considered **unlikely prosecutable**. Finally, cases where all experts label negative are considered **not prosecutable hate speech**. Table 5.1 presents the class distribution.

By examining texts labeled differently by the annotators, differences and similarities in the legal principles are observed. For instance, the Italian hate speech law (Law 167/2017) overlooks gender issues, resulting in most hate speech instances being outside the scope of prosecutability, whereas they might be punishable under the laws of Greece and the UK. On the other hand, there is a gap in interpretation of legal coverage against hateful language towards immigrants in Greek and Italian legislation, where many cases considered negative are deemed positive according to UK law.

5.1.2 Selected PLMs

Using the new legal expert-judged dataset, four BERT pretrained models are fine-tuned and tested. The assessment departs from three hate speech BERT models and one legal BERT model:

HateBERT. (Caselli et al., 2021) based on the English BERT base model (Devlin et al., 2019). It was further trained for the task of hate speech detection on more than 1 M posts from banned communities from Reddit.

DehateBERT. (Aluru et al., 2020) fine-tuned on multilingual BERT[4] and trained for the task of hate speech detection with different learning rates in a monolingual English setting, using various established hate speech datasets.

HateRoBERTa. (Vidgen et al., 2021b) which used a human-and-model-in-the-loop process for training an online hate detection system using RoBERTa (Liu et al., 2019).

LegalBERT. (Chalkidis et al., 2020) which is pretrained on 12 GB of diverse English legal text from several fields (e.g., legislation, court cases, contracts) scraped from publicly available resources, using BERT-base (Devlin et al., 2019).

The assessment of the capabilities of the models is carried out on the basis of Leave One Out Cross Validation (LOOCV). That is, 100 train-and-testing runs are performed in order to have all instances as the test set once. In all cases, the models are fine-tuned for 50 epochs, with early stopping with patience 2, with a learning rate of 1e-5, an AdamW optimizer, and a batch size of 32. The output layer has four units corresponding to the four classes, and softmax is used as the activation function. The classes are mapped accordingly:

$$\text{output} = \begin{cases} \text{non prosecutable} & \text{if class} = 0 \\ \text{unlikely prosecutable} & \text{if class} = 1 \\ \text{likely prosecutable} & \text{if class} = 2 \\ \text{prosecutable} & \text{if class} = 3 \end{cases}$$

The models are evaluated using Mean Absolute Error (MAE) and Mean Squared Error (MSE), with a focus on the differences between the models and the accuracy of their predictions compared to the gold-standard. A lower error value indicates better model performance, with a maximum MAE of 3 and a maximum MSE of 9. The results are described in Section 5.2. Additional F_1 scores are provided in Appendix B.4.

5.1.3 Selected LLMs and Prompting Strategies

Two LLMs and six prompting strategies are also employed to extend the benchmark while investigating possible biases of the LLMs in favor of laws of a specific

[4] https://github.com/google-research/bert/blob/master/multilingual.md

country. The LLMs are Qwen2-7B-Instruct (Yang et al., 2024) and Meta-Llama-3-70B-2-9b (AI@Meta, 2024). Both zero- and few-shot settings are considered:

Zero-shot agnostic (0-shot). The prompt asks to what degree a given instance should be considered prosecutable hate speech. In this case, no example is included (be it positive or negative) from the dataset, nor a legislation, in the prompt.

Zero-shot with law (0-shot w/Law). The same as in 0-shot, but this time the full country-specific hate speech law is also provided in the prompt.

Few-shot agnostic (n-shot). As in 0-shot, but this time a balanced amount of n randomly-selected instances is concatenated to the prompt.

Few-shot with law (n-shot w/Law). As in 0-shot w/Law, but this time a balanced amount of n randomly-selected instances is concatenated to the prompt.

LOOCV agnostic. As in the n-shot, but this time 99 instances selected from the dataset are concatenated, leaving out one instance each time for validation, and use the remaining instances as the prompt.

LOOCV with law. As in the LOOCV, but this time the law is also concatenated to the prompt.

In the few-shot settings, $n \in \{4, 8, 12\}$ are used, as we aim to distribute examples evenly across the four classes whenever possible. Therefore, the distribution is for $n = 4 : 1$ example per class , for $n = 8 : 2$ examples per class, and for $n = 12 : 3$ examples per class, which is the maximum number of examples we could equally take from each class. We chose this balanced setup in order to avoid that the model be nudged more toward classes that are shown more frequently.

5.2 Experimental Results

The experimental results are presented in three segments, beginning with the benchmarking of the PLMs, followed by an analysis of the two LLM-based results, and concluding with a discussion on the impact of silver labels on the performance of the PLMs.

Country	**HateBERT**		**DehateBERT**		**HateRoBERTa**		**LegalBERT**	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Greece	0.60	0.92	**0.54**	**0.68**	0.56	0.88	0.61	0.85
Italy	0.58	1.12	0.48	0.64	0.51	1.01	**0.44**	**0.62**
UK	0.59	1.19	0.59	0.89	0.50	0.98	**0.48**	**0.80**

Table 5.2: MAE and MSE of the PLMs fine-tuned and evaluated with LOOCV. In gray background are the best scores per model. In bold are the best scores per country.

5.2.1 PLM Leave-One-Out Cross-validation

Table 5.2 provides a comparative analysis of the performance of the different models in predicting the ordinal values representing the degrees of hate speech prosecutable in the laws of the three different countries.

DehateBERT consistently demonstrates the lowest error rate for the laws of Greece, achieving the lowest MAE (0.54) and MSE (0.68), indicating its robustness in this context. For Italy, LegalBERT performs best, with the lowest MAE (0.44) and MSE (0.62), showing that it is more adept to handle instances annotated according to the Italian hate speech law. HateRoBERTa performs better with the UK-law-annotated part of the dataset, with the lowest MAE (0.50), while LegalBERT shows the overall best MAE (0.48) and MSE (0.80) for the UK. This variation among the models highlights the importance of evaluating them within specific linguistic, cultural, and legal contexts to ensure their effectiveness.

Overall, LegalBERT demonstrates fewer errors when evaluated against experts' judgments based on hate speech legislation. One possible reason for this is that LegalBERT's pretraining included legal documents, particularly those related to UK legislation. This highlights the potential value of pretraining models on legal texts, when considering the task of prosecutable hate speech, as an alternative to relying solely on datasets of hate speech instances.

5.2.2 LLM Prompting

Tables 5.3a and 5.3b provide insights into the LLMs' performance with different prompting setups. It is noted that in the majority of the cases the error scores drop when we integrate the actual legislation in the prompt. In some cases the performance is improved the more examples we use, as we see with Qwen2, in

Country	**w/o Law**						**w/Law**					
	4-shot		**8-shot**		**12-shot**		**4-shot**		**8-shot**		**12-shot**	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Greece	0.91	1.53	0.89	1.61	**0.87**	**1.47**	**0.78**	**1.22**	0.84	1.42	0.91	1.57
Italy	0.92	1.68	0.88	1.72	**0.67**	**1.27**	0.58	1.98	0.63	0.99	**0.57**	**0.91**
UK	**0.90**	**1.88**	0.94	1.96	0.93	2.11	0.98	2.02	0.91	1.67	**0.73**	**1.29**

(a) Qwen2

Country	**w/o Law**						**w/Law**					
	4-shot		**8-shot**		**12-shot**		**4-shot**		**8-shot**		**12-shot**	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Greece	**0.90**	**1.50**	1.21	2.47	1.02	1.94	**1.06**	**1.88**	1.11	2.23	1.16	2.14
Italy	**1.08**	**2.06**	1.25	2.63	1.35	3.03	**1.08**	**2.04**	1.21	2.47	1.12	2.20
UK	**1.09**	**2.33**	1.15	2.59	1.22	2.66	**1.08**	**2.18**	1.18	2.46	1.15	2.19

(b) Llama3

Table 5.3: MAE and MSE for the few-shot approaches. The column "w/Law" indicates that we also use the actual hate speech law in the prompt. In gray background are the best scores per setting. In bold is the best score per country.

the 12-shot and the LOOCV setups. Although the error scores are low, the overall performance is lacking, as there are cases where one of the classes is frequently completely ignored by the LLM. This does not happen with the multiclass PLM models.

When comparing the two models, Qwen2 outperforms Llama3, which hints that it already contains more legal knowledge incorporated during training. With regard to which law-annotated dataset allows better prediction when incorporated in these models, in the 12-shot and the LOOCV, we see that the best scores are achieved when using the Greek and UK law-annotated datasets, except for the few-shot setting with Qwen2, where the best scores are achieved with the Italian law-annotated datasets. However, the Greek and UK laws are also more comprehensive on a natural language level since they contain more details about the possible targeted minorities and the potential punishable hate speech contexts, as noted by the experts in Section 5.1.1 (see also Appendix B.1).

National Law	**Class: is prosecutable?**			
	Yes	**Likely**	**Unlikely**	**No**
Greece	147	59	47	351
Italy	177	74	50	344
UK	170	79	47	324

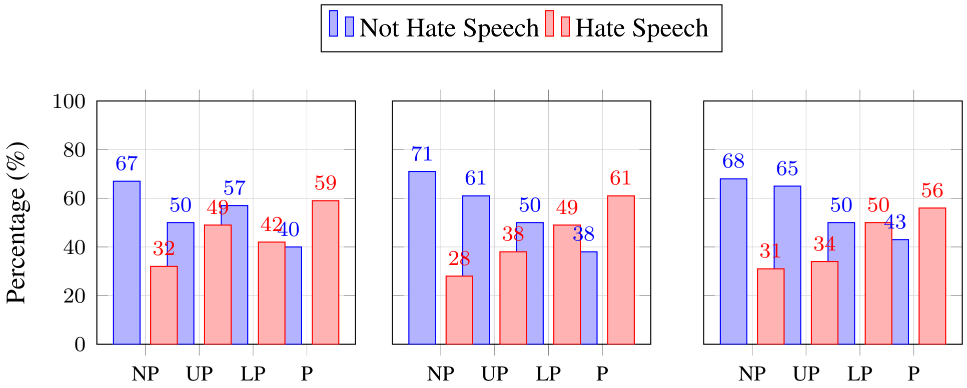
Table 5.4: Prosecutability of Silver Labels under National Laws: Number of cases classified as ‘No’,‘ Potentially’, and ‘Yes’ for prosecution in Greece, Italy, and the UK.

5.2.3 PLMs: Revisited

Following the benchmark with PLMs and LLMs, the following research question is asked: “Can LLM-generated instances be used to improve the performance of PLMs in prosecutable hate speech detection?”. To answer this question, the best performing LLM is used, i.e. Qwen2, to predict on 1,000 instances of the HateEval dataset, creating unseen silver labels. According to Casula et al. (2024), a classifier trained on generated silver labels can lead to a performance that is on a par with, and sometimes even better than, a classifier trained on the original human-annotated data. In this section this intuition is followed, while the possibility of any bias towards any of the country laws the LLM might introduce to the PLMs is also examined.

Figure 5.1 illustrates the distribution of the model’s prosecutable predictions when aligned with the original hate speech labels. Contrary to what was observed during manual annotation, the ‘Not prosecutable’ label is still the most prevalent. However, it is followed by the ‘Prosecutable label’, then ’Likely prosecutable’, and with the ‘Unlikely prosecutable’ label being the least frequent. From this, it can be inferred that there could be a significant number of prosecutable cases for instances that are originally judged simply as hate speech in available hate speech datasets. Examining the instances identified as ‘Not prosecutable’ by the model, reveals notable model inaccuracies, more specifically, a considerable rate of false positives (31% for Greece, 32% for Italy, and 36% for the UK). Analyzing this contrast between hate speech and prosecutable hate speech enables us to filter the silver dataset by eliminating instances of false positives and false negatives. Detailed class statistics for the finalized silver dataset are provided in Table 5.4.

The PLMs are fine-tuned with a similar setup as in Section 5.1 but this time the initial 100 instances are used exclusively as the test set. The results are illustrated in Table 5.5. At a first glance, there is a significant deterioration with regard to predicting the prosecutability of the instances with the pretrained models. Upon manual examination and looking at the per class F_1 scores in the Appendix (Table B.2), it is observed that most models consistently miss to assign the label ‘Likely prosecutable’ or ‘Unlikely prosecutable’, with the majority labeling the instance as ‘not prosecutable’ or ’prosecutable’. This could be due to a variety of factors, including the complexity and ambiguity of the task, the distribution of labels within the dataset, or inherent uncertainties in the input data. Going back to the research question, LLM-generated data do not improve the overall scores.



(a) Greece

(b) Italy

(c) UK

Figure 5.1: Distribution of classes as predicted on 1,000 instances from HateEval (Basile et al., 2019). NP = Not Prosecutable, UP = Unlikely Prosecutable, LP = Likely Prosecutable, and P = Prosecutable. Bars compare hate speech and non-hate speech predictions.

5.2.4 Error Analysis

To wrap up, an error analysis is performed, looking at which classes the errors occurred for each model and setting. The PLMs consistently exhibited low error scores, ranging from 0.50 to 0.60 MAE, with the majority of their errors concentrated in the middle categories (Unlikely prosecutable, Likely prosecutable). In comparison, Qwen2, under its best configuration (12-shot with law), achieved a MAE of 0.57. However, its errors were primarily observed in categories 0 and 1, indicating a shift in error distribution relative to the PLMs. Llama 3, on the other

	Greece		**Italy**		**UK**	
Model	MAE	MSE	MAE	MSE	MAE	MSE
HateBERT	1.30	2.94	1.90	5.04	1.71	4.47
DehateBERT	1.92	4.48	1.96	5.22	1.93	5.13
HateRoBERTa	1.95	4.89	2.10	5.64	2.13	5.73
LegalBERT	1.00	2.06	1.13	2.73	0.79	1.71

Table 5.5: MAE and MSE of the pretrained models fine-tuned on the filtered silver labels and evaluated on our expert-annotated dataset.

hand, showed even higher error rates, with a similar concentration of errors in categories 0 and 1. The confusion matrices of the predictions of each model can be found in Appendix B.5.

Fine-tuning PLMs with silver labels derived from the LLMs was found to increase error rates, suggesting that training with silver data may introduce more noise rather than improving model performance.

Given these results, it appears that pretrained multiclass models specifically designed for hate speech detection are more effective in the context of ordinal regression, as they likely incorporate relevant legal knowledge concerning hate speech.

5.3 Conclusion

This study explored detecting prosecutable hate speech and the effectiveness of introducing legal knowledge during fine-tuning and prompting, using four PLMs and two LLMs. A newly judged sampled dataset on prosecutable hate speech is presented, followed by experiments using the labeled data to prompt them to predict whether a given hate speech instance is prosecutable. Additionally, Qwen2 was used to create silver data and evaluate on the re-annotated dataset on prosecutable hate speech. The results showed that the models struggle to grasp the subtleties of these laws and they often fail to incorporate the specific legal information provided by the laws into their decision-making, relying more on the knowledge they gained during pretraining. The models tend to be more conservative in predicting an instance as prosecutable compared to non-prosecutable. This caution might derive from understanding the legal constraints and requirements with regard to what constitutes prosecutable hate speech and what does not. This tendency mirrors the approach of human annotation, where all experts assigned labels similarly to the models, being generous with the “Not Prosecutable” class. Finally, the error analysis showed that multiclass PLMs trained using LOOCV can be more effective than LLMs for the task of prosecutable hate speech detection.

The next chapter will present the most fine-grained approach to the issue of hate speech definitions by providing a comparative framework that will allow examining differences and similarities of hate speech definitions from different domains (including law) and cultures.

Chapter 6

Untangling Hate Speech Definitions

A Semantic Componential Analysis Across Cultures and Domains

Chapters 2 and 3 introduced the issue of bias and how cultural perspectives influence how hate speech is perceived by individuals. When it comes to annotation, datasets consist of statements produced by individuals within a culture, so the biases reflect, to some extent, the values, norms, and ethics of that culture (Bagga and Piper, 2020; Hershcovich et al., 2022). Since most NLP research focuses on English-language data (Søgaard, 2022), this cultural dimension is often overlooked, resulting in biases that favor English-speaking cultures.

Current NLP approaches are not adequately equipped to address the cultural dependency of hate speech. Existing monolingual hate speech classifiers often lack cultural awareness (Lee et al., 2024). Prevailing hate speech taxonomies tend to focus more on legal or academic definitions rather than incorporating cultural dimensions, a gap that can prove detrimental, as hate speech per se and hate speech regulation might influence societal discourse, relationships, and cultural norms, potentially shaping how people interact and express themselves (Hietanen and Eddebo, 2023b).

Inspired by the compositionality principle (Hinzen et al., 2012), this chapter introduces a component-annotated resource for hate speech definitions, the HateDefCon dataset. The data collection and annotation procedure is described in Section 6.2. A Semantic Componential Analysis (SCA) is proposed, in which a hate speech *definitional component* is defined as a fundamental element or criterion referring to what constitutes hate speech in terms of target, intention/purpose, and act/means. More on SCA can be found in Section 6.1. As *definitional hate speech domains* are defined the contexts where hate speech definitions emerge. This study focuses on five such domains: legislation, Wikipedia, online dictionaries, research papers, and conduct policies from online platforms or technological companies. This resource and framework also allows the examination of the cultural representation of hate speech definitions. The term *culture* is used as the term that encompasses language, ideas, beliefs, customs, codes, institutions, tools, techniques, among other elements.[1] This chapter highlights the need for cross-cultural and cross-domain approaches to define hate speech and argues that such definitions should be context-specific, be it cultural, legal, or academic, grounding hate speech definitions within these particular domains. This is in line with best practices for tackling subjective tasks (Rottger et al., 2022), in which the guideline - or the definition - chosen should consider the downstream use, i.e. the context, as it is mentioned in Chapter 2.

This chapter focuses on three research questions:

1. What are the differences among various definitions of hate speech?

2. What is the diversity of these definitions?

3. How do definitions with different components affect LLM predictions?

The contributions are both theoretical and practical. On the theoretical front, the cross-cultural and cross-domain analysis of definitions shows significant variation in components, ranging from broad definitions to highly specific ones (see Section 6.3) . Even among the more detailed definitions, which address aspects like the target of hate speech, the intent, and the methods of expressing it, there are differences in their components. On the practical side, this chapter (specifically Section 6.4) delves into the assessment of whether LLMs respond better to certain definitions, potentially revealing underlying biases. The results (Section 6.5) reveal that definitions vary in their components, while domains also borrow definitions from one another. When research is culturally specific, borrowing from other domains can be problematic, as it may introduce elements that do not align with the intended cultural context.

A comparable study to this one is the one by Khurana et al. (2022), who develop hate speech criteria with input from legal and social science perspectives to help researchers create precise definitions and annotation guidelines. They propose five criteria: target groups, dominance, perpetrator characteristics, type of negative group reference, and potential consequences/effects. Rather than prescribing a single definition, they offer a meta-prescriptive, modular approach, allowing

[1] https://www.britannica.com/topic/culture

for adjustments based on specific tasks. This approach emphasizes the role of subjectivity in definitions and addresses the issues arising from varying and vague definitions, which can lead to inconsistencies and problematic expectations about dataset annotations. The framework introduced in the present chapter is a more fine-grained one, and it elaborates on the impact of the components when applying different definitions when prompting LLMs.

6.1 Semantic Componential Analysis

Semantic Componential Analysis (SCA) is a linguistic technique used to break down the meanings of words or phrases into their constituent parts or features. This method, central to structural linguistics, has been in use since the 1950s (Lounsbury, 1956; Goodenough, 1956; Nida, 1975). It is based on the *principle of compositionality*, which states that the meaning of a complex expression is derived from the meanings of its components and the rules governing their combination (Hinzen et al., 2012). SCA involves compiling a detailed list of specific examples for each term within a group of contrasting terms. Each example is described using a set of relevant attribute dimensions (Kronenfeld, 2005). SCA primarily examines words through organized sets of semantic features, which are marked as 'present' or 'absent', using +/- symbols (Geeraerts, 2006).

Table 6.1 illustrates an example of SCA application.

	Parent	Sibling	Male	Female
Father	+	-	+	-
Mother	+	-	-	+
Brother	-	+	+	-
Sister	-	+	-	+

Table 6.1: Kinship terms characterized by attribute dimensions. The attribute dimensions include Parent, Sibling, Male, Female. Each example is marked with ’+’ if the feature is present, ’-’ if absent.

SCA is an approach that enables us not only to break down terms into their individual components, but also to explore various types of meanings as categorized by Leech (1990). By comparing terms side by side, we can enhance our analysis, going beyond the conceptual meaning (the stable meaning across contexts),

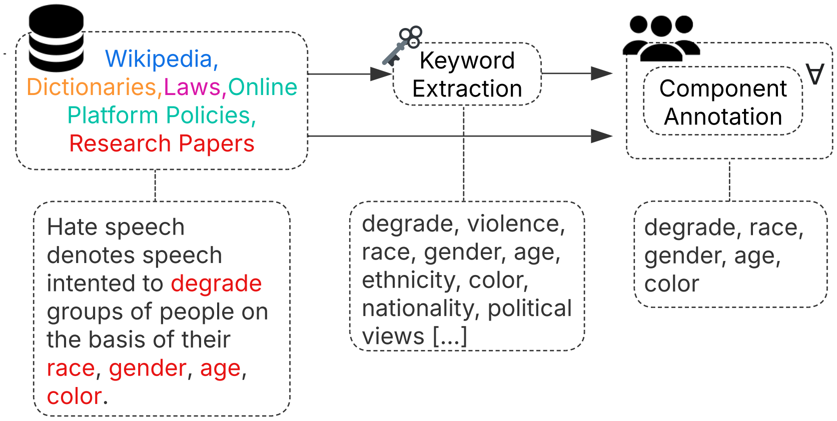


Figure 6.1: `HateDefCon` creation pipeline.

touching upon the connotative and social meanings. These latter meanings can vary depending on cultural and social contexts.

Kronenfeld (2005) describes the identification of components. An approach involves systematically alternating attributes to determine which distinctions are essential to differentiate terms. Another method is to gather descriptions from informants about differences between terms or subsets of terms, gradually building a set of potential semantic components based on these descriptions. The methodology presented in this chapter draws inspiration from the latter approach.

6.2 The `HateDefCon` Dataset

In this section, hate speech definitions are analyzed and compared via SCA to identify potential cross-domain and cross-cultural differences. The definitions are collected from five different domains in which we can find hate speech definitions.

Figure 6.1 summarizes the `HateDefCon` dataset creation pipeline. The definitions are collected from the selected domains, and then keywords are extracted in order to be used as components for each definition during annotation. More details are found in the following sections.

6.2.1 Data

Hate Speech Laws. The definitions are sourced from the Global Handbook of Hate Speech Laws website, a comprehensive resource that provides access to existing hate speech legislation from countries worldwide, including the United Nations and European Union levels.[2] The legislations are already available in English. We exclude countries for which their legislative texts were not available. The full list of the countries is available in Appendix C.1.

Wikipedia. The domain is Wikipedia articles in different languages from which the relevant portions of the definitions are manually extracted. The extracted portions are automatically translated into English.[3]

Dictionaries. The second domain is dictionaries across multiple languages. As with the Wikipedia definitions, the extracted content is translated into English using machine translation.

NLP Research Papers. For this domain, the Hate Speech Dataset Catalogue[4] is used to navigate through research articles and datasets, focusing specifically on definitions of hate speech. Other related terms like toxicity or abusive language are not included, as the primary focus of this study is hate speech.

Online Conduct Policies. The last domain is conduct policies from online platforms or technology companies that address the use of hate speech or harmful language online. The conduct policies were collected from X, Meta, Microsoft, Pinterest, Snapchat, YouTube, Reddit, TikTok, and Discord.

6.2.2 Component Annotation

Three annotators are engaged for SCA labeling. Annotators are proficient in English, have experience in annotating tasks, and are familiar with hate speech research. All three of them annotate the entire corpus. They are provided with detailed instructions on annotating definitions following our framework, as outlined in Appendix C.2. To compile an initial set of components, keyword extraction

[2] https://futurefreespeech.org/global-handbook-on-hate-speech-laws/.
[3] https://www.deepl.com/en/translator.
[4] https://hatespeechdata.com/

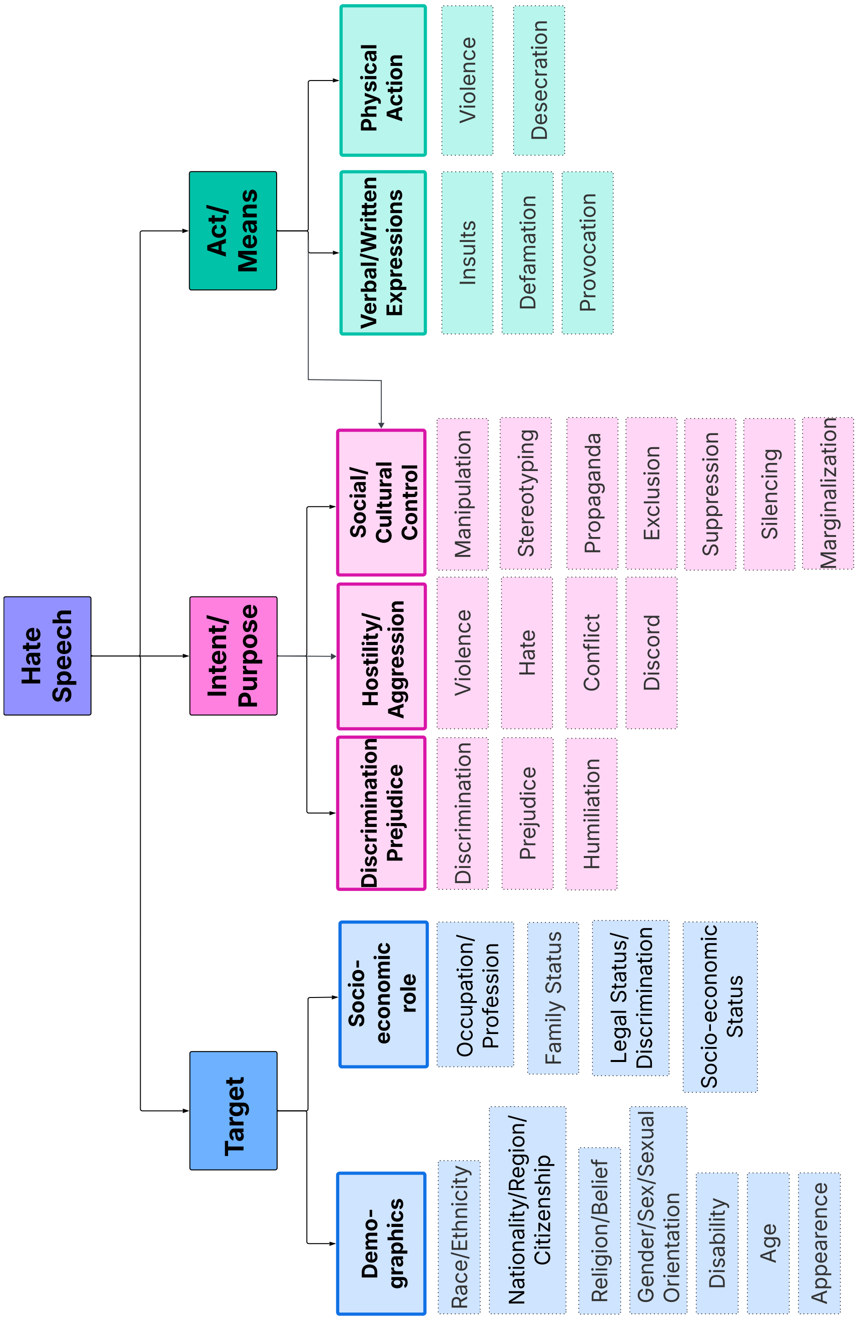


Figure 6.2: SCA component hierarchy in HateDefCon.

is employed and tf-idf on the collected definitions. A manual review reveals that these automated methods fail to capture some components, highlighting the necessity for human annotation. Despite this limitation, these computational techniques offer a valuable starting point. Using the resulting list, annotators then perform binary annotations, indicating the presence or absence of each component.

To ensure objectivity in the annotation process and a very fine-grained representation of the components, it is established that a definition contains a given component if a derivative of the component's word appears in the annotation. For instance, if the word ‘abusive’ is present in the definition, the annotator marks the component ‘abuse’ as present. A challenge we encounter during annotation is the treatment of synonyms. For instance, the words ‘harm’ and ‘hurt’ can be considered synonymous, but they are not derivatives of the same root word. To maintain consistency and avoid subjective interpretation, which could lead to different grouping of the synonyms by the annotators, only annotate derivatives of the target term are annotated (e.g., marking ‘harm’ only if words like ‘harmful’ or ‘harmed’ appear). Additionally, missing components are labeled as ‘undefined component’. This process allows to comprehensively gather all components deemed crucial by the annotators.

To finalize the dataset, all annotated components are kept, as a manual inspection showed that most of the disagreement derived from the fact that one of the annotators missed a component. The average IAA measured via Cohen's kappa is 0.64. Figure 6.2 reports the component hierarchy resulting from the annotation study (see Appendix C.3 for more details on the IAA and Appendix C.4 for a fine-grained version of the hierarchy).

6.3 Case Study: Definition Comparison

Qualitative Analysis. A qualitative assessment of the definitions is performed, which provides some initial insights into the data. The initial observations reveal that many definitional domains, particularly in research papers and Wikipedia articles, frequently borrow definitions from other sources. For instance, 17 out of 30 research papers reference definitions from other academic papers, legislation, or platform policies. Wikipedia, on the other hand, often relies on definitions from the Cambridge Dictionary. Through translation, definitions might reflect the culture of the source text, and in collaborative projects, cultural elements from multiple backgrounds may become blended.

Domain	**Definitions**	**Culture Distribution**	**Components**
Law	116	All cultures appear once	Race, Religion, Hate, Nationality, Ethnicity
Wikipedia	49	All cultures appear once	Religion, Gender, Sexual Orientation, Race, Disability
Research Paper	29	English (13), German (3), Arabic (2), Indonesian (2), the rest appear once	Gender, Religion, Sexual Orientation, Ethnicity, Race
Dictionary	21	English (7), Italian (5), the rest appear once	Attack, Gender, Sexual Orientation, Religion, Hate
Platform	278	All cultures appear once per platform	Disability, Ethnicity, Gender, Nationality, Race
Total	493		

Table 6.2: Number of definitions, cultural distribution, and top 5 components per domain. Some domains correspond to languages (e.g., Wikipedia, dictionaries, platforms), others to countries (e.g., laws), or remain unspecified (e.g., some research papers). Consequently, cultures encompass all possible variations.

Distributions. Table 6.2 reports culture distribution for each domain in HateDefCon. A great disparity is evident with the English definitions, which appear most often in research papers and dictionaries, while platform policies are always the same text but translated in different languages. Definitions from other cultures appear fewer than 5 times, and most occur only once in the dataset. Table 6.2 also reports the 5 most frequent components per domain. The most common components are related to the target of hate speech, mainly being associated with religion, ethnicity, and gender.

Cross-Cultural Case Study. SCA is employed to describe potential inconsistencies between collected hate speech definitions and the cultural reality through a real case study. Mulki et al. (2019) develop a Levantine Hate Speech and Abusive Twitter dataset in an attempt to bring into the spotlight less-spoken Arabic varieties. However, in their study, the authors refer to the hate speech definition by Nockleby (2000), which defines hate speech as “any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic”. Levantine Arabic

Components	Mulki et al. (2019)	Israel Legislation	Syria Legislation	Jordan Legislation
Target				
Demographics	Ethnicity, Religion, Sexual Orientation, Color, Gender, Nationality, Race	Ethnicity, Religion, Race, Color	Race	Ethnicity, Religion, Race
Intent				
Discrimination	Disparage	Humiliation, Degradation	Race	Prejudice
Hostility	N/A	Hostility, Enmity	N/A	Hurt, Hate, Conflict, Violence
Cultural Control	N/A	Persecution, Degradation	N/A	N/A
Act				
Expression	Disparage	Hostility, Humiliation, Enmity	N/A	Hate
Physical Action	N/A	N/A	N/A	Hurt, Conflict, Violence, Terrorism
Manipulation	N/A	Persecution, Cultural Control	N/A	N/A

Table 6.3: Comparative analysis of components of Mulki et al. (2019) vs. the legislation of countries where Levantine Arabic is spoken in. N/A means that the component is unavailable.

is spoken in many Middle Eastern countries, such as Syria, Jordan, and Israel. SCA allows us to see if the definitions available for these countries overlap with the one used in Mulki et al. (2019) (Table 6.3). While they all differ in the intent and act, some targets overlap, although the definition from Mulki et al. (2019) is more comprehensive, considering also sexual orientation and gender, which are not taken into account by the other definitions. Israel's legislation seems to have the most detailed provisions, including concepts of hostility and cultural control, whereas Syria's legislation is narrowly focused on race, and Jordan's legislation emphasizes physical action and terrorism. This variation can affect annotation and prompting, resulting in varied interpretations of hate-related content. Therefore, tailoring prompts to align with specific regional definitions is essential for achieving consistent model behavior for a specific culture.

6.4 LLM Sensitivity to Definitions

This section evaluates the ability of LLMs to perform the task of binary hate speech classification based on the definitions provided. In particular, this section examines the reliance of the LLMs on the definitions of hate speech provided rather than their own internal knowledge.

Models and Data. The experiments are conducted with three open-source popular state-of-the-art LLMs: Llama3[5], Mistral2[6], and Phi3-mini.[7] The data used are the Gab Hate Corpus (GHC) on online hate speech conversations (Kennedy et al., 2022). GHC consists of 27,665 posts from the social network service gab.com and is annotated for the presence of "hate-based rhetoric" by a minimum of three annotators. The posts are annotated based on a coding typology created by synthesizing definitions of hate speech from legal precedents, existing hate speech coding frameworks, and insights from psychology and sociology. This typology includes hierarchical labels that denote dehumanizing and violent speech, as well as indicators related to targeted groups and rhetorical framing. This dataset captures various dimensions of hate speech, making it well-suited for a detailed and fine-grained analysis and testing with multiple definitions. Although GHC is annotated in a multi-class classification and since the definitions of hate speech that might vary in components are examined, a binary setting is considered since : hate vs not hate. To limit computational overhead, 500 instances are selected from the corpus, equally divided between the two classes.

Setup. The experiments are conducted in a zero-shot setting to understand the impact of different definitions (i.e. different components) on model performance without any additional fine-tuning or prompting strategies. To prompt the models, three definitions of hate speech are used. The definitions are selected deliberately after manual inspection, aiming to assess whether the varying degrees of component coverage influence the prompting outcomes. This approach includes one definition with medium component coverage, a highly detailed one, and a very general one. Three definitions are used: $\mathtt{D}_{ghc}$ comes from the GHC dataset, it includes several components but is not the most comprehensive one; $\mathtt{D}_{wiki}$ is the definition provided by the Macedonian Wikipedia page, it offers a more detailed and varied

[5] `Llama3-8B-instruct`.

[6] `Mistral-7B-Instruct-v0.2`.

[7] `Phi-3-mini-4k-instruct`.

Model	D_{ghc}	D_{wiki}	D_{dict}
Llama3	0.67	0.78	0.70
Mistral2	0.56	0.53	0.58
Phi3-mini	0.60	0.58	0.58

Table 6.4: Classification performance (F_1 score) using difference hate speech definitions.

perspective; and D_{dict} is a general definition from the Merriam-Webster dictionary (see Appendix C.5 for more details). The temperature is set to zero to exclude variations in the generated response.

6.5 Results

Table 6.4 summarizes the results of the classification task according to the three definitions. Llama3 performs consistently better than other models. The complexity of the definitions affects the outputs of the three models differently. In Llama3, the comprehensiveness of the definition is directly proportional to the performance: the more comprehensive the definition is, the more the performance increases; in Mistral2, a reverse tendency is observed. Phi3-mini outperforms Mistral2 in all settings, although being half the size of Mistral2. Additionally, in Phi3-mini, we observe the smallest variety of responses based on the definitions, with D_{ghc} performing the best.

There are cases in which Llama3 refuses to answer for safeguarding the generation process over harmful content. This tendency is also proportional to the completeness of the definition. The less comprehensive the definition, the more the model refuses to answer it. For D_{wiki}, the model refuses to answer to 1% of times, 1.6% for D_{ghc}, and 5% for D_{dict}.

Error Analysis. An error analysis of the answers generated by Llama3 is also performed. The error analysis encompasses two aspects: how the model perform considering the definitions against the gold labels, and how the predictions differ according to the definition used. Figure 6.3 shows the confusion matrices with respect to the three definitions. In all three settings, the false negatives can partly be attributed to the lack of components in the definition. For instance, D_{ghc} focuses on violence and hate, excluding all other forms of hate speech. As a result,

True \Predicted	0	1
0	226	24
1	103	147

(a) $\mathtt{D}_{ghc}$

True \Predicted	0	1
0	214	36
1	67	183

(b) $\mathtt{D}_{wiki}$

True \Predicted	0	1
0	222	28
1	102	148

(c) $\mathtt{D}_{dict}$

Figure 6.3: LLama3 confusion matrices with respect to each hate speech definition. When the model refuses to answer, we set the predicted label to zero.

discrimination against a group of people that does not exhibit violence or hate often goes undetected. However, the presence or absence of such components is not a definitive factor. Take the following example sentence: *"Real friends don't let their friends get infected with communism"*. The sentence is incorrectly categorized as non-hateful by $\mathtt{D}_{wiki}$, even though it contains political views as a component. While not conclusive, analyzing the components of misclassified examples can contribute to understanding the reasons behind the classifications.

How do definitions affect the output? All instances for which the answer of Llama3 varies according to the definition are extracted, regardless of the ground truth. $\mathtt{D}_{ghc}$ and $\mathtt{D}_{wiki}$ differ in 54 instances, while $\mathtt{D}_{dict}$ differs from $\mathtt{D}_{ghc}$ and $\mathtt{D}_{wiki}$ in 48 instances each. We observe that when the model is prompted with $\mathtt{D}_{wiki}$, it tends to identify more personal attacks, which are often discarded in $\mathtt{D}_{ghc}$ and $\mathtt{D}_{dict}$ as false negatives since such definitions identify group of people as targets of HS, rather than individuals. Moreover, with $\mathtt{D}_{wiki}$, LLMs tend to identify more instances that are hateful with respect to political views and undocumented migrants. Indeed, the majority of instances that are correctly identified by $\mathtt{D}_{wiki}$ and misclassified by $\mathtt{D}_{ghc}$ and $\mathtt{D}_{dict}$, contain terms like liberals, communists, and illegal aliens with a negative connotation.

Cross-Cultural Analysis. We go back to the case study of the Levantine dataset (Table 6.3) to explore how these definitions affect prompting and whether

Def.	$\texttt{D}_{lev}$	$\texttt{D}_{isr}$	$\texttt{D}_{syr}$	$\texttt{D}_{jor}$
$\texttt{D}_{lev}$	1.00	0.74	0.59	0.50
$\texttt{D}_{isr}$	-	1.00	**0.75**	0.67
$\texttt{D}_{syr}$	-	-	1.00	0.50

Table 6.5: Agreement across definitions via F_1 score.

Def.	$\mathbf{\texttt{D}}_{isr}$		$\mathbf{\texttt{D}}_{syr}$		$\mathbf{\texttt{D}}_{jor}$	
	P	**N**	**P**	**N**	**P**	**N**
$\texttt{D}_{lev}$	55	309	55	295	55	339
$\texttt{D}_{isr}$	-	-	190	288	157	299
$\texttt{D}_{syr}$	-	-	-	-	166	294

Table 6.6: Number of overlapping instances in the cross-cultural setting with separate positive (P) and negative (N) values.

cultural biases may arise as a consequence. A prediction comparison is performed for the definition from the original Levantine dataset (Mulki et al., 2019) ($\texttt{D}_{lev}$), the one for Syria ($\texttt{D}_{syr}$), the one for Israel ($\texttt{D}_{isr}$), and the one for Jordan ($\texttt{D}_{jor}$). The four definitions are in Appendix C.6. Llama3 is considered for the analysis as the best-performing model. Llama3 achieves the top performance (F_1=0.74) with $\texttt{D}_{isr}$ and $\texttt{D}_{syr}$, exceeding most of the results with the original definition. This is an interesting result since $\texttt{D}_{syr}$ is the least comprehensive definition, with only one target component: *race*. In contrast, using $\texttt{D}_{lev}$ leads to significantly lower scores (F_1=0.32). Lastly, with $\texttt{D}_{jor}$ Llama3 achieves F_1=0.67, falling behind best results, yet still largely outperforming its variant using $\texttt{D}_{lev}$. F_1 score across the four definitions is also computed. Each time, it is assumed that one definition represents the gold standard when compared to another one. Table 6.5 shows the results. The highest F_1 (0.75) is reached between $\texttt{D}_{syr}$ and $\texttt{D}_{isr}$, followed by $\texttt{D}_{lev}$ and $\texttt{D}_{isr}$, with a small 0.01 point difference. However, in terms of components, the definitions used in Israel and Jordan are the most similar, as they share three target components: ethnicity, religion, and race. Table 6.6 shows the number of overlapping instances across the four definitions, divided into positive (P) and negative (N). The agreement between $\texttt{D}_{syr}$ and $\texttt{D}_{isr}$ is confirmed by the number of overlapping instances, as they have the greatest agreement with 478 instances. In all cases, most of the overlapping instances are on the negative class, as it is the class that the model tends to overpredict.

6.6 Conclusion

This chapter introduced `HateDefCon`, a comprehensive hate speech definition dataset. `HateDefCon` provides detailed component annotations, capturing the target, intent/purpose, and act/means of collected hate speech instances that allow comparing hate speech definitions. The analysis of `HateDefCon`, revealed a lack of cultural diversity in existing definitions. This is primarily because only legislative sources referred to specific cultures. Wikipedia definitions are also often the result of collective contributions or translations of the original English text, making them less culturally distinct. There is also a variation in terms of components, especially when one language can refer to multiple cultures, such as in the case of Levantine Arabic. Moreover, the experiments showed that LLMs are sensitive to the employed hate speech definitions, where, in some cases, more comprehensive definitions lead to better results.

This work underscores two key considerations for hate speech detection research: (a) Definitions to be incorporated into the annotation guidelines or prompts must be specific to the task and should clarify the level of comprehensiveness or generality required for that task. This consideration should also align with model selection, as some models perform better with general definitions while others with more comprehensive ones. (b) Definitions must be relevant to the language and the target culture. This may involve referring to hate speech legislation, to understand what constitutes hate speech in the given culture, otherwise clarify that no culture is considered.

The next chapter will continue to address the development of resources for hate speech detection, with a particular emphasis on tackling the challenges of multilinguality and the challenges it entails.

Chapter 7

A Parallel Multilingual Hate Speech Corpus

An Exploration of Challenges

Both Chapters 5 and 6 discussed how accommodating linguistic and cultural diversity in hate speech annotation and automatic detection is overlooked, regardless of definitional context. While some multilingual and cross-lingual approaches exist (Lee et al., 2023a,b; Arora et al., 2023), there is still a need for multicultural and cross-cultural approaches, as well (Hershcovich et al., 2022). Language- and culture-sensitive approaches should begin at the level of the annotation and corpus creation, as biases in supervised models are integrated from the very first step. However, employing different annotators that are proficient in different languages or from diverse cultural backgrounds can be a difficult and expensive task.

In this chapter, the focus is on cross-linguality, while also touching upon cross-culturality. In an attempt to enrich the parallel resources that could be used for hate speech detection purposes, a pipeline is designed that allows filtering online hate speech instances that can be translated with a minimum effect on meaning and toxicity levels of the original text. Specifically, examples from an already existing hate speech dataset (Davidson et al., 2017a) are extracted and used to automatically generate translations into Greek and Italian. Then, backtranslation (Ueffing et al., 2007) is employed in order to perform a quality check of the translation (Moon et al., 2020). A BERT-based toxicity classifier (Devlin et al., 2019) is applied in the original, translated, and backtranslated sentences to produce toxicity scores. Finally, a qualitative analysis on the translations is performed so as to identify any patterns that could help in the optimization of the pipeline.

The rest of the chapter is structured as follows. In Section 7.1, the method is

Name		Language	Instances
Hate Speech/Offensiveness (Davidson et al., 2017b)		EN	1,000
Offensive Greek Tweet (Pitenis et al., 2020)	(FT)	EL	4,779
HaSpeeDe@EVALITA (Sanguinetti et al., 2020)	(FT)	IT	6,837
Measuring Hate Speech Sachdeva et al. (2022a)	(FT)	EN	5,966

Table 7.1: Statistics for the data used in the translation experiments, as well as the data used for fine-tuning Mbert (FT). This table includes the language, the total number of used instances.

described, including the data and the model that were used. In Section 7.2, the results are shown, followed by a qualitative analysis in Section 7.3. Finally, a discussion is provided in Section 7.4, along with the conclusions and future steps in Section 7.5.

As defined by Barrón-Cedeño et al. (2015), parallel texts are essentially precise translations or close approximations with only slight language-specific differences when compared to a comparable corpus, which should ideally consist of texts in multiple languages that are similar in both structure and content. Therefore, the term 'parallel multilingual data' is used to refer to the dataset.

7.1 Methodology

The proposed methodology can serve as a means of identifying high-quality translation instances, which could potentially be incorporated into a parallel corpus, without requiring human experts. In this way, not only will the cost be reduced as we rely less on human evaluation, but also the process will be streamlined, allowing for the creation of additional parallel corpora for addressing hate speech detection and other NLP tasks.

Data The data used for the translation experiments are instances extracted from the Davidson et al. (2017a) dataset, which originally contained instances of both hate speech and offensiveness. For the purposes of this study, only those that are labeled as hate speech are extracted, as the focus is on this linguistic phenomenon, touching upon cultural implications. This is also due to the fact that the primary objective of this study is to examine the challenges of the parallelization of hate speech, and not hate speech detection per se.

For fine-tuning our model, the Offensive Greek Tweet dataset is used (Pitenis et al., 2020) consisting of offensive and non-offensive text samples, and an Italian hate speech dataset, HaSpeeDe@EVALITA2020 (Sanguinetti et al., 2020), which consists of hate speech and non-hate speech instances. For English, a sample of the Measuring Hate Speech dataset (Sachdeva et al., 2022a) is used, which also contains hate speech and non-hate speech instances. Table 7.1 shows basic statistics of the data.

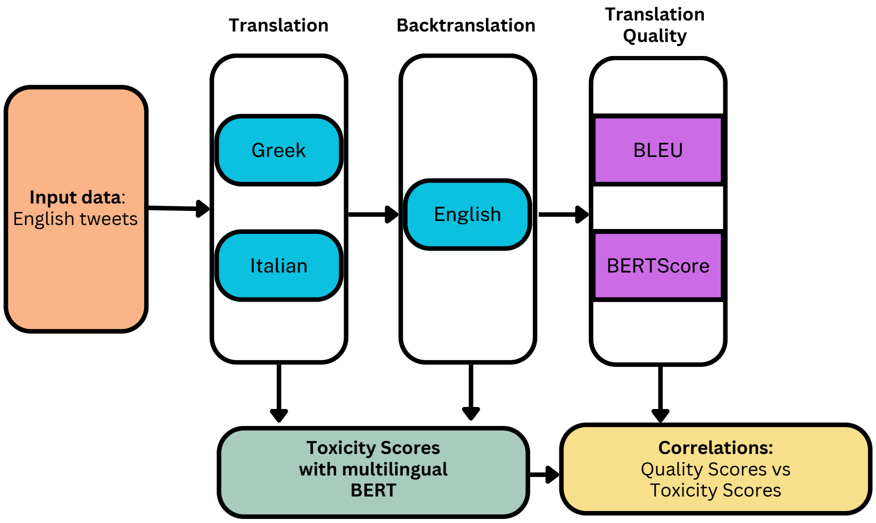

Translation
Backtranslation
Translation Quality
Input data: English tweets
Greek
Italian
English
BLEU
BERTScore
Toxicity Scores with multilingual BERT
Correlations: Quality Scores vs Toxicity Scores

Figure 7.1: Visualization of translation and evaluation pipeline.

Models The model experiments predominantly revolve around machine translation, therefore the out-of-the-box translation engine of ModernMT is used (Bertoldi et al., 2018) through their official API[1]. It is a context-aware, incremental and distributed general purpose Neural Machine Translation system based on the Fairseq Transformer model (Ott et al., 2019). To evaluate the translation quality, two metrics are employed; the first one is BLEU (Papineni et al., 2002), in order to see the similarity of n-grams and evaluate the translations on a structural level. The second

[1] `https://www.modernmt.com`; translations were carried out between September and October 2023

is BERTscore (Zhang et al., 2020), which allows us to examine the translations on a semantic level. BERTscore measures the similarity between embeddings, thus it is used as a proxy to assess the quality of the translation, while also taking into account semantic and contextual information.

In terms of toxicity scoring, multilingual BERT (mBERT) is used (Devlin et al., 2019) by training three separate monolingual models on the English, Greek and Italian data, while also training one single multilingual model with all the data simultaneously. More specifically, the models are initialized with the 'bert-base-multilingual-cased' checkpoint and are further trained on the Offensive Greek Tweet dataset (Pitenis et al., 2020), on the HaSpeeDe@EVALITA2020 dataset (Sanguinetti et al., 2020), and the Measuring Hate Speech dataset (Sachdeva et al., 2022a). During fine-tuning, AdamW optimizer is used with a learning rate of 2e-5 and trained the model for three epochs. The input text is tokenized with a maximum sequence length of 128, which was the default.

Pipeline The research method revolves around producing translations with a minimum change in the original meaning and toxicity levels with respect to the source text. To achieve this while avoiding the need for human experts or manual translators, and to automate the process as much as possible, a round-trip translation is opted for, which means producing translations and backtranslations (Lee et al., 2023b; Beddiar et al., 2021). Figure 7.1 presents the pipeline.

First, a pilot experiment is conducted with a limited dataset of 100 instances, manually checking the quality of the translation, as well as generating quality and toxicity scores, as intended in our main experiments. Since good translation quality from ModernMT is observed, a larger-scale experiment is rendered feasible by translating and backtranslating the whole set of 1000 hate speech instances. Once obtained all the translations and the backtranslations, BLEU and BERTScore are computed on the backtranslations as the hypothesis and with the source text as the reference.

Then, the fine-tuned mBERT is applied on the original English text, Greek and Italian translations, and the backtranslations to produce toxicity scores for all versions. In addition, accounting for the possibility that toxicity scoring might not capture deeper semantic meaning, a qualitative analysis is also conducted, to see on which level we can compromise by using only computational methods.

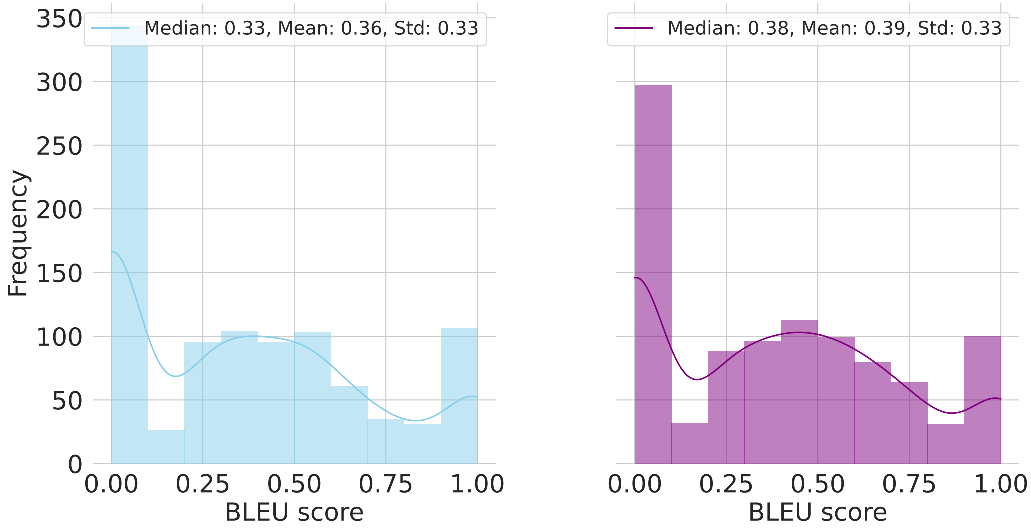


Figure 7.2: Distribution of BLEU scores for backtranslations from Greek (left) and from Italian (right).

7.2 Experimental Results

In this section, the experimental results of the translation quality (Section 7.2.1) and toxicity evaluation (Section 7.2.2) are presented. The translation quality is assessed by comparing the translations back into the source languages with the original source texts. To measure translation quality, BLEU score and BERTScore are used. Additionally, toxicity levels are analyzed in the source text and translations, including their backtranslations, using a BERT-based toxicity classifier.

7.2.1 Translation Quality

In order to asses the translation quality, the backtranslation into the source language is compared with the original source texts. Figure 7.2 shows the distribution of BLEU scores on instances back-translated both from Greek and Italian. Overall, BLEU scores are relatively low. While approximately 200 instances achieve scores between 0.60 and 1.00, indicating high quality, the majority of scores fall below this range. In fact, over 300 instances have a BLEU score of 0. The backtranslations from Italian slightly outperform those from Greek, with more instances surpassing the threshold of 0.6.

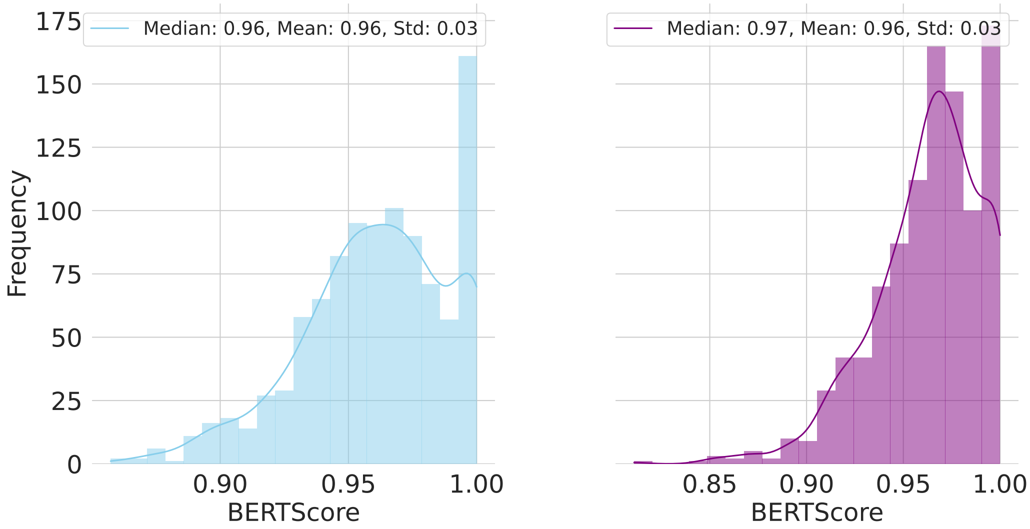


Figure 7.3: Distribution of BERTScores for backtranslations from Greek (left) and from Italian (right).

Figure 7.3 shows the results in terms of BERTScore. Both Greek and Italian scores are quite high, with the values falling within the range of 0.8 to 1.0, which indicates an adequate match between candidate backtranslation and reference.

This indicates that some translations may not be very accurate when evaluated with the BLEU metric, suggesting a limited n-gram overlap. In contrast, the BERTScore values for both Greek and Italian backtranslations look more promising. Considering that in this work we value semantic similarity more than the structural sentence matching, the translations appear to be better in terms of capturing semantic content.

7.2.2 Toxicity Evaluation

The primary objective of the toxicity evaluation is to assess whether the toxicity is retained throughout the entire pipeline procedure. The averaged toxicity scores are presented in Figures 7.4 for individually trained monolingual and 7.5 for multilingual models.

When scoring with the mBERT models that are trained separately in the three languages, we observe some discrepancies. The English texts, including both the source and the backtranslations, have the higher averaged toxicity score, ranging

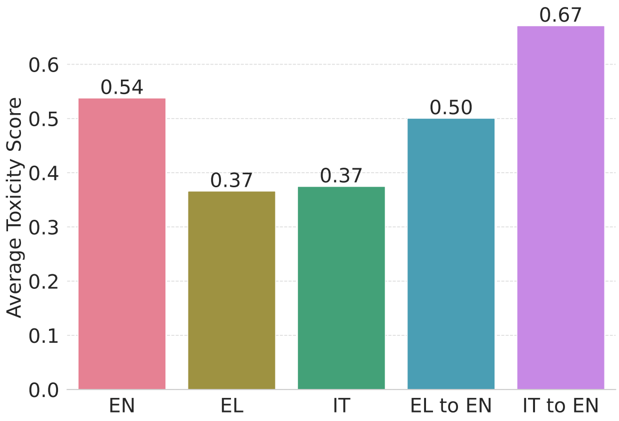


Figure 7.4: Average toxicity scores across monolingual models. The **EN** bar represents instances in the original language (English), **EL** corresponds to translations into Greek, and **IT** represents translations into Italian. **EL to EN** indicates backtranslations from Greek to English, while **IT to EN** represents backtranslations from Italian to English.

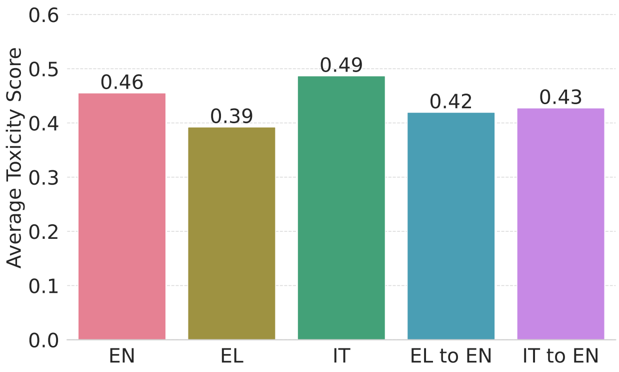


Figure 7.5: Average toxicity scores from multilingual models. The **EN** bar represents instances in the original language (English), **EL** corresponds to translations into Greek, and **IT** represents translations into Italian. **EL to EN** indicates backtranslations from Greek to English, while **IT to EN** represents backtranslations from Italian to English.

from 0.54 to 0.67, while the score is lower for the Greek and Italian translations to 0.37. However, we would have expected more pronounced discrepancies in the toxicity levels of the English texts, indicating that the reason might be that backtranslations restore the toxicity of the text. In fact, the toxicity when backtranslating from Italian to English exceeds the toxicity of the initial input texts. This could be due to semantic shift that might occur during translation (Beinborn and Choenni, 2020), which leads to change or even amplification of certain meanings of the text. Similarly, the multilingual model scores the toxicity of the Italian translation as more toxic compared to the original English text.

Overall, when examining the toxicity with the multilingual model, there is less fluctuation among translations and backtranslations compared to the monolingual models, with the values ranging from 0.39 to 0.49. Still, the toxicity scores of the backtranslations are lower than the toxicity scores of the original text, hinting that indeed some toxicity is lost during the translation process.

Yet, we must take into account that potential errors in machine translation could influence the toxicity score in subsequent phases of the pipeline. Hence, a qualitative assessment is also conducted to address this concern (see Section 7.3).

7.3 Quality Evaluation

7.3.1 Threshold Filtering

To explore the possibility of filtering the sentences and keeping those that have maintained their toxicity, as well as assessing whether they are adequately translated, several thresholds are defined that allow filtering a number of sentences and manually evaluate them in terms of meaning and toxicity, using the scores produced from the multilingual model. More specifically, the steps that are followed for the filtering and the manual evaluation are:

1. Calculate the absolute difference between the source sentence's toxicity score (T_{src}) and the toxicity score of the sentence translated into Greek/Italian (T_{trans}):
$$\Delta_{\text{trans}} = |T_{\text{src}} - T_{\text{trans}}|$$

2. Calculate the absolute difference between the source sentence's toxicity score (T_{src}) and the toxicity score of the backtranslated sentence from Greek/Italian (T_{back}):
$$\Delta_{\text{back}} = |T_{\text{src}} - T_{\text{back}}|$$

Threshold	Number of Instances	Description
0.1	141	Instances with unchanged toxicity scores
0.2	333	Instances with unchanged toxicity scores
0.3	611	Majority of the dataset with unchanged toxicity scores

Table 7.2: Number of instances retaining toxicity scores at different thresholds.

3. If both absolute differences are less than or equal to a specified threshold τ, the sentence is added to the list of maintained sentences:

$$\text{If } \Delta_{\text{trans}} \leq \tau \text{ and } \Delta_{\text{back}} \leq \tau, \text{ then include the sentence.}$$

Using a threshold of **0.1**, 141 instances that maintained their toxicity scores are identified. When the threshold is increased to **0.2**, 333 sentences that meet the criteria for maintained toxicity scores are identified. Further increasing the threshold to **0.3** results in the identification of 611 instances that continue to meet the criteria for maintained toxicity scores, and is in fact more than half of the initial dataset. The 0.2 threshold is kept, as setting the threshold too low might result in capturing instances with maintained toxicity but poor translation quality. A threshold of 0.2 allows a compromise between quality and toxicity maintenance. A small sample of these instances is then qualitatively evaluated.

7.3.2 Manual Evaluation

From the 333 filtered instances, 60 instances are randomly selected for further evaluation by two language experts. One is proficient in both English and Greek, holding a linguistics degree in both languages, and possesses Greek as their mother tongue. Likewise, the language expert who evaluated the translations from English to Italian holds a degree in Languages, with Italian as their mother tongue. In this section, different linguistic and cultural features of some of the borderline cases are discussed, as well as some of the successfully translated parallel sentences.

One of the initial observations in both languages is the overall semantic quality of the translations, as corroborated by the results of our quality control analysis. On the contrary, grammatical and syntactical inconsistencies create more certain complexities in achieving the parallelization of the sentences. We present and analyze some noteworthy examples in Table 7.3.

Example 1 is one of the cases where translation was successful conveying the intended meaning but generated grammatical issues. In particular, in the Greek

			BERTScore	
Source English	**Translation Greek**	**Translation Italian**	**Greek**	**Italian**
1. kill yourself ugly ass monkey looking shit	αυτοκτονήστε άσχημη μαϊμού που δείχνει σκατά	uccidi te stesso brutto culo scimmia cercando merda	0.96	0.97
2. all us beaners were pissed like robert when the jew ref stopped it	όλοι εμείς οι φασολιάρηδες τσατιστήκαμε σαν τον Ρόμπερτ όταν ο Εβραίος διαιτητής το σταμάτησε	tutti noi beaners eravamo incazzati come Robert quando l'arbitro ebreo lo fermò	0.96	0.96
3. shut the fuck up fag ill beat the shit outta you	βγάλε τον σκασμό αδερφάρα να σε πλακώσει στο ξύλο	chiudi quella cazzo di bocca da frocio ti ha pestato a sangue	0.93	0.96
4. shut your dirty dirt trap you limp knees, ashy footed, blue balls, isolated hamster wheel, no nipples bitch	κλείστε τη βρώμικη παγίδα σας, χαλαρά γόνατα, στάχτη, μπλε μπάλες, απομονωμένη ρόδα χάμστερ, χωρίς ϑηλές σκύλα	chiudi la tua sporca trappola per sporcizia, ginocchia zoppicanti, piedi cenere, palle blu, ruota isolata per criceti, niente capezzoli, stronza	0.94	0.94
5. Fucking gook	Γαμημένε κιτρινιάρη	Fottuto muso giallo	1.00	0.92

Table 7.3: Noteworthy translation cases for the qualitative analysis.

translation there is an incorrect use of the person in the main verb ('kill'), which disagrees with the subject 'ugly ass monkey'. In Italian, on the other hand, the problem arises with the phrase 'looking shit', which refers to the appearance which is erroneously translated as 'cercando', meaning 'to look for'. *Example 2*, in contrast, is one of the successful cases of translation where both the meaning and the grammatical structure of the sentence are preserved in both Greek and Italian. In Greek, the word 'beaner' which is a racist slur towards Mexican people, is translated as φασολιάριδες which is a Greek adaptation that sounds natural. In Italian, however, the word beaner remains unchanged which is acceptable, as English words sometimes are integrated intact into other languages, especially in modern usage.

Example 3 has also managed to capture the intended meaning of the original instance in both target languages, yet there is a shared error in terms of the person. This error might possibly be due to 'ill' which could be texting language for 'i'll', and which the models falsely substituted with the third person in both languages. *Example 4* is one of those examples where the words were translated accurately yet the meaning changed because of the literal aspect of the translations. More specifically, in the source text, the author of the tweet uses a series of objectifying insults to make up a misogynistic comment. These sort of insults are common in neither Greek nor Italian, and therefore the translations do not sound natural. Finally, *Example 5* is similar to Example 1 where the translation was successful for both Greek and Italian. The specific example is a racial slur, mainly towards Asian people, therefore, in both target languages, the translation referred to the color 'yellow' (κιτρινιάρη, muso giallo) which is usually associated with East Asian people, thus capturing the toxicity of the original text.

7.4 Discussion

The findings of this study, as well as previous research endeavors (Lee et al., 2023b,a) show that the creation of a parallel hate speech corpus is feasible, however there must be a degree of compromise to overcome several obstacles.

Challenge 1: The quality of the translation is not certified. In Section 7.2.1, we observe that the performance of the machine translation system is adequate. However, although the semantic evaluation yielded high scores, there were serious grammatical and syntactical issues with the sentences, leading to lower BLEU scores. Creating a comprehensive parallel corpus demands both semantic and grammatical/structural correctness. As explained by Hershcovich et al. (2022), linguistic form and style are associated with social and cultural factors, and any linguistic variations must be correctly represented in datasets. Therefore, both choosing the right translation method and the right metrics for the evaluation are paramount to ensure quality. Given this, the method presented in this chapter allows us to filter out higher quality sentences.

Challenge 2: Toxicity fluctuates from source to target language. Only preserving the intended meaning is not sufficient when creating a parallel hate speech dataset; the levels of the toxicity of the original text must also be maintained. This is not an easy task because, as we saw in Section 7.2.2, the toxicity fluctuates when translating to another language. Therefore, an individual study should be conducted on the toxicity levels in order to analyze toxicity fluctuations to ensure that, beyond meaning and sentence structure, toxicity remains on similar levels of that of the original text.

Challenge 3: There is culturally embedded meaning that is hard to be translated. As has been mentioned throughout this thesis, a major issue in NLP and hate speech detection revolves around the lack of cultural awareness. The qualitative analysis was an attempt to shed some light on the cultural dimensions of hate speech. The analysis showed that there are instances that were effectively translated, yet there might be missing cultural context. In our case, Example 2, which employs the derogatory term 'beaner' in reference to Mexican people, is adequately translated, yet it resonates primarily with a specific geographical population, namely individuals from the US. People from other countries and cultures may struggle to comprehend why this specific example constitutes hate speech. This is an example of the task of *adaptation* in translation, as described in Peskov et al. (2021). The authors assert that, while computational techniques for this task have advanced, there is still room for improvement. They still advocate for the automatizing of the task and they recommend using available datasets in high resource languages by adapting content instead of just translating it literally. Our approach paves the way for this endeavor.

Taking all these challenges into account, it is clear why creating a parallel hate speech corpus is not an easy task. In this study, the amount of sentences that were filtered and could be compiled in a parallel dataset were limited but they existed, rendering the task hard but still feasible. Especially, since machine translation is becoming more and more effective, it should be "used for bridging between cultures, investigating cross-cultural communication" (Hershcovich et al., 2022).

7.5 Conclusion

This chapter addressed the challenges of creating a parallel multilingual hate speech corpus. The methodology involved utilizing machine translation to translate English tweets into two target languages—Greek and Italian—and then employing backtranslation along with translation metrics to assess the quality of these translations. Additional evaluations were conducted to analyze the toxicity levels of the translated texts, culminating in a qualitative review of a sample of the instances used. The findings revealed that while machine translation can adequately convey the intended meaning of a sentence, it often introduces grammatical and syntactical errors that render the output unsuitable for inclusion in a parallel corpus. Only a limited number of examples successfully preserved both the meaning and the toxicity of the original text without such issues, something that underscores the difficulty of the task.

The following chapter adopts a psycholinguistic perspective to investigate harmful language, emphasizing how various textual elements affect the annotation process.

Chapter 8

Psycholinguistic Effects

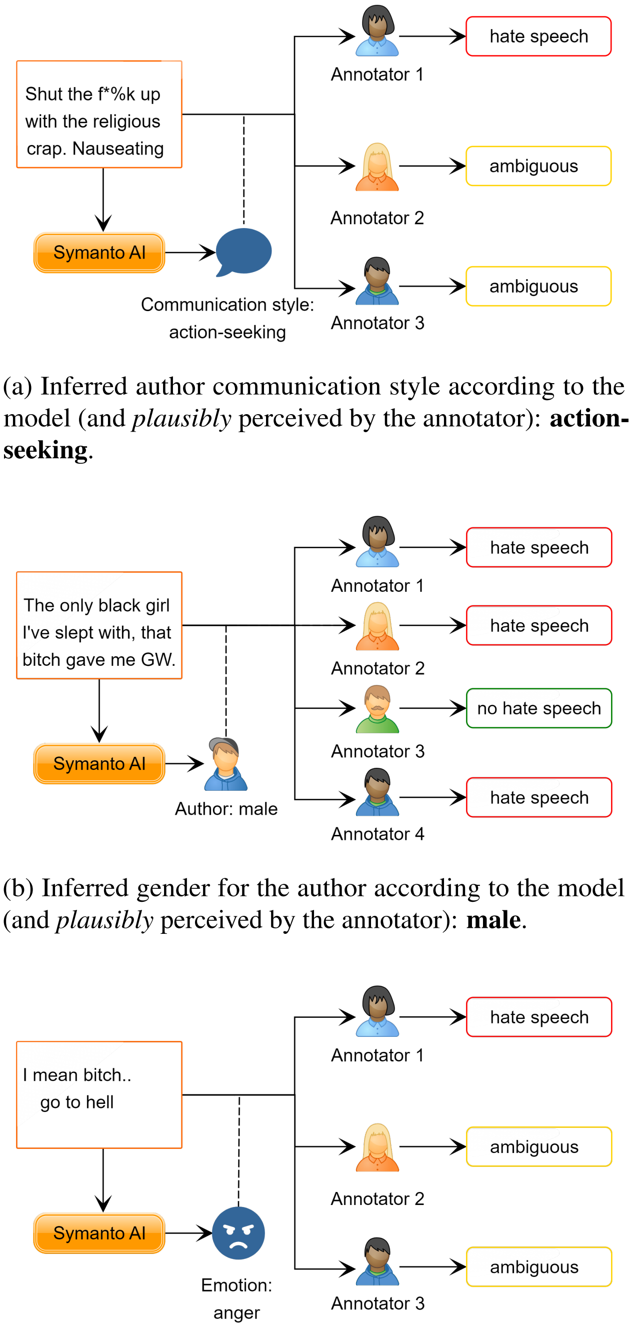
Examining Inferred Author and Textual Correlates of Harmful Language Annotation

In Chapter 3, we saw that how annotators label instances can depend on many factors such as the annotators' personal experiences and beliefs, which can induce their own personal biases in AI models and/or the implicitness of the text itself (Goyal et al., 2022; Sap et al., 2022; Ocampo et al., 2023). The problem of annotator bias is a crucial one as it can be "ethically and socially concerning, when [it] can reinforce algorithmic oppression against minoritized or marginalized populations" (Sachdeva et al., 2022b). When AI systems discriminate against certain groups of people, it raises ethical concerns about fairness, justice, and human rights, consequently damaging the reputation and trust of organizations that deploy them.

Apart from the annotators' characteristics, there are factors that lie in the text itself and can influence annotation and create problems with model evaluation. These factors include language features such as sarcasm and implicitness (Frenda, 2018), as well as ambiguity and variation (Beck et al., 2020). Similarly, several factors related to the communicator (in the case of this thesis, the author) have been identified as predictors of how their communication style is perceived by the recipient (Von Thun, 1981; Leung and Bond, 2001). Thus, exploring how different communication styles can be associated with hate speech annotation is a crucial point of this research. Figure 8.1 presents three examples that allow the formulation of the hypothesis that text aspects and inferred author demographic information, such as the communication style, gender, and emotion can be correlated to how annotators label harmful language. As shown in Figure 8.1a, one of the annotators labels the sentence as hate speech, while the other two find it neutral or ambiguous.

This may be the result of a misinterpretation of the sentence or due to the action-seeking tone conveyed through the use of the imperative mode which may have led to confusion, as the definition of hate speech is usually associated with incitement of hatred and violence (Jahan and Oussalah, 2023), where the use of the imperative is commonly used. Particularly, this instance cannot be classified as hate speech as, although it contains the imperative mode, it does not incite any violence towards some marginalized group, as state most hate speech definitions (Hietanen and Eddebo, 2023a). In addition, studies show that language use and wording vary depending on the author's demographic information, such as gender and age (Pennebaker and Stone, 2003; Flekova et al., 2016), which can in turn lead to different text interpretations. With that in mind, in this chapter the role of the author's gender and age in annotation is also explored. Figure 8.1b showcases the variation in annotations for a sentence predicted to be authored by a man. Notably, the two female annotators classified the sentence as hate speech, whereas only one of the male annotators made the same classification. This forms the hypothesis that the perception of hate speech may be influenced by text features that convey author-related information, such as gender, emotion, sentiment, or communication style. I propose that these textual aspects could correlate with how annotators perceive and classify hate speech. There are also cases where the emotion potentially plays a role in the annotation of harmful language. An example is presented in Figure 8.1c. While "bitch" is often not classified as hate speech in a general sense, it can be misogynistic or offensive depending on context (Pinker, 2007). In this case, the overall expressed emotion of anger may lead the annotators to perceive such language as more hostile or aggressive, increasing the likelihood of classifying it as hate speech.

As language cues can *betray* different psycholinguistic attributes, I examine communication style, sentiment, emotions and emotionality of the text itself as another factor that can bias the annotators' perspective during harmful language annotation, as emotionally charged language and style of communication can lead to subjective interpretations of harmful language. More specifically, in this study, leveraging psycholinguistic textual elements serve as a proxy to discern whether author attitudes can indeed influence annotation, shedding light on the intricate dynamics at play among author, text, and annotator during the annotation process.



(a) Inferred author communication style according to the model (and *plausibly* perceived by the annotator): **action-seeking**.

(b) Inferred gender for the author according to the model (and *plausibly* perceived by the annotator): **male**.

(c) Inferred emotion of the author according to the model (and *plausibly* perceived by the annotator): **anger**.

Figure 8.1: Instances of annotated hate speech with annotator information from the Measuring Hate Speech dataset (Kennedy et al., 2020)

The aim of this work is to analyze annotator and text characteristics in relation to bias in harmful language detection datasets. In a holistic attempt, I examine certain text aspects —psycholinguistic information, namely sentiment, emotions, emotionality and communication style, as well as inferred demographics of the author(s) of the text, such as gender and age— to see how they might affect the perception of harmful language, specifically 'toxicity' and 'hate speech', during annotation. I examine the characteristics of the text in relation to the annotations by fitting a statistical regression model to data from two different datasets: the *Jigsaw Specialized Rater Pool* (SRP) dataset (Goyal et al., 2022) and the *Measuring Hate Speech* dataset (MHS) (Kennedy et al., 2020). The results show that most text aspects are correlated with labeling harmful language.

This chapter seeks to find answers to the following research questions:

RQ1 How are psycholinguistic aspects of the text (emotion, emotionality, communication style, sentiment) associated with labeling for hate speech and toxicity purposes?

RQ2 Do certain inferred demographic aspects of the author of the text (in particular, age and gender) affect the annotators' perception when labeling for hate speech and toxicity?

The rest of the paper is structured in the following manner: In Section 8.1, I discuss some social psychology and psycholinguistic approaches to harmful language and its relation to annotator characteristics. I present our methodology and the reasons behind our selected data analytics and models, followed by a brief overview of the computational models used for our analysis in Section 8.2. I present our results and analysis in Section 8.3 followed by a discussion in relation to our research questions in Section 8.4. Finally, I discuss the limitations of our study in Section 8.5.

8.1 Psycholinguistic Characteristics of the Text

As in most annotation tasks, in hate speech and toxicity detection annotators, most often, do not have any informed knowledge about the authors of a potentially toxic or hateful text and, therefore, making assumptions regarding the intention of the authors of a text is inevitable (Fortuna et al., 2022; Sap et al., 2019). Thus, one of the primary aspects of toxicity and hate speech is its perception from the target groups who get exposed to and observe or are victims of hateful acts. The perception of target groups varies depending on internal factors that are inherent such as age, gender, cultural background, and ideology, and contextual factors expanding to the social experiences such as prior exposure and being a member of a disadvantaged group (Sap et al., 2022; Chetty and Alathur, 2018; Guberman and Hemphill, 2017; Cowan and Hodge, 1996). For example, Guberman and Hemphill (2017) show that female annotators are better at detecting more subtle forms of aggression than men, possibly because women are more likely to have personally experienced certain types of microaggressions. All these factors play a significant role in triggering certain emotional responses (e.g. fear, anger, sadness, outrage) associated with perceived threat, and can lead target groups to identify hate speech and toxic language in particular ways (Boeckmann and Liew, 2002).

Among the attributes that could influence annotators' perception of harmful language, this chapter focuses on the psycholinguistic features of the text they encounter and rate in terms of toxicity and hate speech. Psycholinguistic markers of a text give better insights than the content itself and manifest various psychological states of an individual (Pennebaker and King, 1999). Several studies revealed that it is the case for toxic language as well. Specifically, psychological states and conditions of the hate speech perpetrators spill over to their language, which could be tracked as personality variables (e.g. right-wing authoritarianism), emotional states (e.g., disgust, hate), and motivations (e.g., thrill-seeking) (Duckitt, 2001; Gerstenfeld, 2002; Cottrell and Neuberg, 2005; McDevitt et al., 2002). Humans are fairly good at observing and perceiving such psychological and personality traits since making judgments is a natural part of day-to-day interactions, and lots of short- and long-term decisions are made through this way (Macrae and Quadflieg, 2010). Research shows that even first and quick impressions provide accurate observations in terms of predicting traits (Carney et al., 2007). In this study, I attempt light on how different types of psycholinguistic text aspects influence annotators' perception of harmful language.

Previous research has shown that comments inclined towards toxicity and hate speech had significantly different linguistic features than non-toxic comments. Social media comments with hate speech were longer, combined with more exclamation and question marks, and contained less numbers of emojis (Gevers et al., 2022). Studies yielded that toxic comments conveyed less agreeable, less emotionally aware and more informal communication qualities (ElSherief et al., 2018a,b). Studies with an emotion focus showed that toxic comments consisted of more anxious, depressed, angry and less joyful cues (Cheng et al., 2019; Chatzakou

Sentiment	**Example**
Negative	Just awful, why can't a-holes like this just kill themselves.
Positive	Oh, lord, he's got himself so worked-up that he's developed a plan!

Table 8.1: Examples of predicted sentiment from SRP.

et al., 2017). Such research evidence supports that hate speech is characterized by particular psycholoinguistic patterns representing author's intentions, desires, emotions and personality traits and likely to influence annotators' perception (ElSherief et al., 2018b; Balayn et al., 2021). In this psycholinguistic examination, four aspects are considered:

DEMOGRAPHICS: Basic demographics have been extensively exploited for linguistic research as they are able to represent personality and psychological traits (Schwartz et al., 2013). Pennebaker and Stone (2003) have found striking psycholinguistic differences and variations predicted by age and gender. For age, they show that as people get older, they use more positive, less negative and less self-disclosed wording. Women use wording about social interactions and contextual features whereas men refer more to formal, informational and affirmative cues (Newman et al., 2008). Somewhat contrary to these findings, a more recent research by Hilte et al. (2023) shows that men produce more hateful comments than women, and people produce more hate speech the older they grow, as well as that specific age and gender dynamics vary slightly in different languages or cultures. Combining this background with the research evidence that people perceive situations and conversations differently when they collect cues about the gender, age and political viewpoint of the author (i.e., categorical thinking), a certain level of bias is expected in this study as well (Quinn and Macrae, 2005; Flekova et al., 2016). Accordingly, we could expect that inferred demographic qualities of the author of a text could impact how annotators perceive its toxicity level.

SENTIMENT: Sentiment analysis is the classification of texts as positive, negative, or neutral, which in turn allows the understanding of emotional dispositions towards topics and situations (Munezero et al., 2014). Studies on sentiment provide insights into the relationship between computational models and psychological measurements, as using the correct computational framework to evaluate a psychological state can aid AI model the human mind (Jo et al., 2017). A line of

Emotion	**Example**
Fear	“as if every Democrat had to state for the record they are not a member of the Communist party.” Or every Muslim to state for the record that they are appalled by terrorism?
Sadness	did you forget the Dallas shooting during the BLM rally. How many police officers died? Was President Obama your President during that time?
Joy	Marriage_is_the_joining_of_families. _When_families_of_gays_want_religious_weddings,_they-will_get_them.
Surprise	oh really? During Fox News coverage of DNC convention, When Khizr Khan spoke, Fox went to a commercial about Bengazi, and returned to Katy Perry singing. What a perfect world you live in!
Anger	I think this is a terrible idea and not because he carries a lot of baggage. Dude was not that accurate, we already have 3 QBs that aren’t accurate. Sure he can scramble but with our line that is all he will be able to do. Add in the locker room, management, and fan division by having him here and pop goes the weasel!
Disgust	"unbeknownst to them, they walked into a bakery owned by a person harboring such prejudice against same-sex couples he refused to bake them a cake. "Complete and utter BS. They were singled out because of their religious beliefs.
No emotion	I did; she may have had black ancestors - but probably not.

Table 8.2: Examples of emotions adapted from SRP.

research suggests that negative sentiment attracts more attention, as people tend to care about conflict, controversy and bias in text (Chmiel et al., 2011; Mejova et al., 2014). On the other hand, positive sentiment is shown to boost user interest and create more virality online than negative content (Berger and Milkman, 2013). In this study, I am interested in understanding how negative and positive sentiment would be associated with the annotation task. Examples of sentiment predictions can be found in Table 8.1.

EMOTIONS: In many cases, sentiment analysis is insufficient and therefore emotion detection studies are also required, which are able to determine an individual’s emotional or mental state (Nandwani and ., 2021). Basic emotions are the most fundamental and distilled versions of emotional experiences that represent the rawest discrete and automatic responses to stimuli, situations and events (Plutchik, 2001; Ekman and Cordaro, 2011). As the most established and studied emotion theory, Ekman's model lays out six basic emotions: *anger, disgust, fear, joy, sadness*, and *surprise* (Ekman and Friesen, 1969). As the basic emotions discern the most prototypical elements of an experience, it is applicable and observable in many contexts. For instance, people who hold prejudice and hate towards minority groups have difficulty in regulating their emotions and tend to be expressive about them (Walters et al., 2016). Harmful language and emotion co-exist in close proximity, as many studies have tried to reinforce hate speech detection by including emotion information. For example, taking into account previous psychological and empirical investigations that associate negative emotions and hate speech states, Min et al. (2023) use multi-task learning, where any hate speech sample can be assigned with two kinds of labels, including a hateful label and an emotion label leading to more accurate hateful predictions. Considering the contribution of emotions in the perception of hostile behaviors, it is important to break down the kind of emotions that impact the annotators' labeling of toxic language. Examples of emotion can be found in Table 8.2.

Emotionality	**Example**
Rational	Cal came out hotter than a firecracker but you're right that it did become hard to watch. If that debacle didn't bring the Ducks back to earth nothing will and now they need to keep , as they say, Cal from beating us twice.
Emotional	Alpha blowhard, perhaps, but haven't we gotten beyond the idea of masculinity as being necessarily the loudest, most arrogant, most selfish, and most obnoxious personality in the room?

Table 8.3: Examples of emotionality adapted from SRP.

EMOTIONALITY: Emotionality in text gives away clues about how people experience the world and explains their way of reacting and coping with situations and events (Tausczik and Pennebaker, 2010). *Emotional* qualities refer to feeling-focused cues involving values and emotions while statements that are less emotionally-charged (i.e., emotionally-neutral) are characterized by *rationality*, which provides logical and analytical reasoning (Štajner et al., 2021). Previous psycholinguistic studies showed that emotional text conveyed increased levels of immersion to an important event (e.g. personal trauma), leading to high levels of emotional processing of the situation (Holmes et al., 2007). On the other hand, text

Communication Style	**Example**
Self-revealing	My understanding is that a PPS employee/fundamentalist Christian complained because Frida is bisexual and it is mentioned in the film.
Fact-oriented	children are mostly born with male or female genitalia. that makes them boy or girl. very few have that strange condition of both. it is this perverted culture that lives to confuse the most innocent. shame on these perverts and purveyors of filth.
Action-seeking	Dear Legislators: Go read the UofA article about people losing their jobs. Get to work on the budget issues.
Information-seeking	Why don't heterosexuals feel the need to parade their heterosexuality annually?

Table 8.4: Examples of predicted communication styles adapted from SRP.

that contained rational cues, such as causal wording, focused more on concrete information about a topic and provided a cognitive mechanism that triggered an active processing and reevaluation of the event (Tausczik and Pennebaker, 2010; Pennebaker et al., 1997). Considering all this, it is debatable whether annotators could be more impacted by emotional textual cues which convey an emotionally-charged tonality or by the rational text due to paying attention to logical linguistic markers in their perception of harmful language. Emotionality examples can be found in Table 8.3.

COMMUNICATION STYLE: The way people communicate showcases their psychological states, leading the receiving end of the communication to have a certain impression about the sender. By sending information and signals in a particular way, the person denotes an appearance, identity and interpretative suggestion to him/herself (de Vries et al., 2013). Multiple variables have been reported about the communicator as predictors of how their communication style is perceived by the recipient. For instance, a friendly style generally predicts attraction and likeability (Leung and Bond, 2001). According to the Four-sides Communication Model (Von Thun, 1981), communication is a dynamic and layered act where every utterance contains messages about facts, self-disclosed information, social context, and desires. By basing upon this communication framework, four different communication styles are used: *self revealing*, statements with personal information and experiences; *fact-oriented*, statements about facts, logic and objective observations; *action-seeking*, requests, demands and suggestions asking for action; *information-seeking*, direct or indirect questions (Štajner et al., 2021). With this, the aim is to understand how different styles of communication would relate to the annotation of hateful language. Examples of each communication style can be found in Table 8.4.

8.2 Proposed Approach

For the purposes of this study, two datasets are selected, one for toxicity detection and one for hate speech detection. Running the models of text aspects, enabled obtaining scores on different demographic and psycholinguistic aspects of the text. At this point, it must be noted that, since this study is the first to examine text aspects in relation to toxicity and hate speech annotation, a ground-truth dataset was not available and, thus, proprietary AI models were used to generate said information. The used AI models are properly evaluated, excel at processing vast amounts of data and can identify complex patterns that might be challenging for humans to detect. An ordinal regression analysis was then performed to find statistical effects and study the relation between the labeling and text aspects. Figure 8.2 illustrates the approach. The first step is using data from two datasets: MHS dataset and SRP. dataset. Instances from these two datasets are inserted into predictive artificial intelligence models that allow to obtain predicted labels on demographics and specific psycholinguistic text aspects. For the final step of this study, ordinal regression is performed with the newly obtained predicted labels as the dependent variable and the toxicity or hate speech scores that were initially included in the two datasets. More information about the newly added text information and the statistical models is described in the next sections. It must also be noted that, for the present study and in relation to the demographic information, binary gender (man-woman) is modeled following the common practice in NLP in order to also make the data and results comparable to other studies, as well as because it is easier for the models to distinguish subtleties between these two genders.

8.2.1 Datasets

Information about the datasets can be found in Table 8.5. The first one is part of the Civil Comments dataset (Borkan et al., 2019b) and is known as the Jigsaw

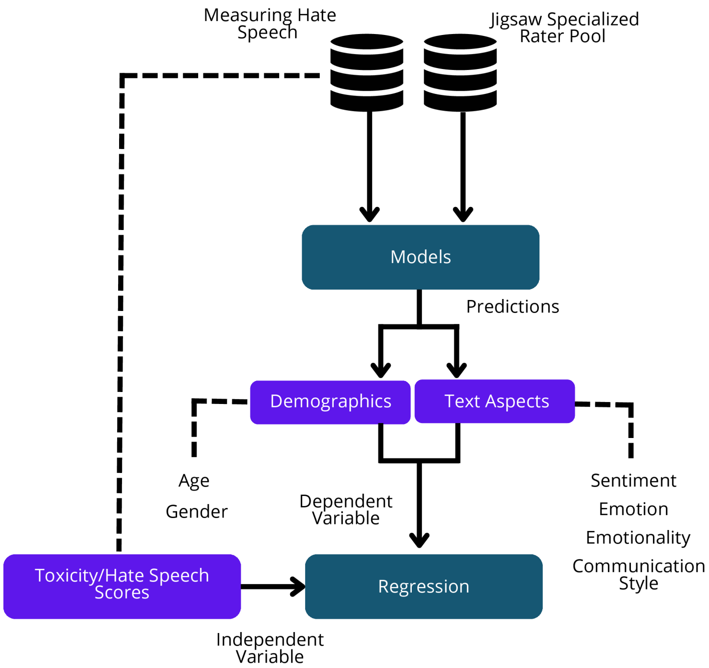


Figure 8.2: Chart that illustrates the selected approach for the purposes of this study.

Specialized Rater Pool Dataset (SRP) (Goyal et al., 2022). The Civil Comments dataset includes crowd-sourced labels for toxicity and subtypes, with 22% of the comments labeled for identity references. Goyal et al. (2022) explore the overall amount of disagreement between rater pools (African American, LGBTQ, Control Group) by showing the percentage of comments where the absolute value of the mean difference is greater than or equal to 1. They find that toxicity has the largest proportion of comments with disagreement (over 12% for both African American and LGBTQ rater pools), whereas threat and profanity show the least amount of disagreement. The rate of toxicity was controlled using Perspective API (Jigsaw, 2017), aiming to mitigate exposure to harmful content for annotators. A total of 25,500 comments were sampled: 8,500 identity-neutral, 8,500 mentioning LGBTQ, and 8,500 mentioning African American identities. This resulted in a dataset with 382,500 individual annotations

	Jigsaw Dataset	*Measuring Hate Speech Dataset*
Number of texts	25,500	50,070
Number of unique annotators	953	11,143
Number of annotations	382,500	135,556
Source	Civil Comments	YouTube, Twitter, and Reddit
Class labels	-2 = Very toxic -1 = Toxic 0 = Unsure 1 = Not toxic	-1 = Counter or supportive speech 0 = Neutral or ambiguous 1 = Hate speech

Table 8.5: Statistical and class information about the two datasets

The second is the Measuring Hate Speech (MHS) dataset (Kennedy et al., 2020). Each comment is assigned 10 ordinal labels, covering aspects such as sentiment, disrespect, insult, attack/defense, humiliation, status (inferior/superior), dehumanization, violence, genocide, and a 3-point label for hate speech. These labels are combined using faceted Rasch measurement theory (RMT) to produce a continuous score that indicates the comment's position on a hate speech spectrum. The study evaluated annotator agreement using Krippendorff's alpha, finding generally weak agreement across all survey items, with alphas below 0.5. The demographic information included in MHS involves the gender of the annotators, sexual orientation, religion, education level, income level, race, and political ideology.

8.2.2 Experimental Setup

Models The Symanto proprietary text analytics API is used for the inferred demographics (age and gender) and psycholinguistic aspects (emotionality, communication style, emotions, and sentiment) of each text. It must be noted that the Symanto API exposes the endpoints upon request.[1]

These API analytics models were trained (before the present study was conducted) via deep learning technology (LeCun et al., 2015) on millions of texts. An annotation study with psycholinguistic experts was also conducted, which led to the creation of a new dataset that was used to train and evaluate the proprietary models on the different communication styles and emotionality (Štajner et al., 2021). Table 8.6 shows statistics for the data partitions used to train the models.

[1] https://api.symanto.net/docs

Aspect	Training		Test	
	NO	YES	NO	YES
Emotionality	6,557	6,538	496	504
Fact-oriented	11,047	2,097	827	173
Self-revealing	2,548	13,162	160	840
Action-seeking	15,969	1,945	884	116
Information-seeking	17,259	3,177	844	156

Table 8.6: Statistics of the dataset partition for the Symanto proprietary models for communication style (Štajner et al., 2021).

Model	Emotionality			Fact-oriented			Self-revealing			Action-seeking			Info-seeking		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
c-CNN	89	89	89	89	80	83	86	86	86	85	81	83	94	93	94
ng-SVM	90	90	90	90	87	88	89	86	87	91	81	85	94	88	91
BERT	95	95	95	94	95	95	94	96	95	92	87	90	96	96	96

Table 8.7: Macro-averaged performances (in %) of the proprietary models used in Štajner et al. (2021).

Performance metrics such as Precision (P), Recall (R) and F_1 score (F), evaluated on the human-annotated test set can be found in Table 8.7. In addition, these models have been benchmarked against state-of-the-art models and have shown similar performance when compared to public and reference benchmarks (Štajner et al., 2021; Chinea-Rios et al., 2022; Mueller et al., 2022; Basile et al., 2021), with the advantage of offering a ready-to-use, stable and fast analytics API.

Once all the values from the different models are obtained, the ordinal regression analysis is performed using an R Package for Bayesian Multilevel Models Using Stan (**brms**) which enables the investigation of the results using various methods defined on the fitted model object (Bürkner, 2017).

8.3 Results

8.3.1 *Preliminary Style Exploration*

Initially, an examination was conducted to determine whether the statement in the psychology literature asserting that social media comments containing hate

	SRP		*MHS*	
	Hate	**No Hate**	**Hate**	**No Hate**
Avg. no. of tokens	82	84	29	31
Avg. no. of characters	414	425	142	155
Avg. no. of exclamation marks	0.31	0.20	0.70	0.22
Avg. no. of question marks	0.66	0.58	0.17	0.19
Avg. no. of emojis	10	10	4	4

Table 8.8: Preliminary exploration, specifically on the differences in sentence length, punctuation, and emoji usage between hate and non-hate content within the two datasets, SRP and MHS.

speech are longer, include more exclamation and question marks, and feature fewer emojis (Gevers et al., 2022) holds true. To that end, the number of tokens, characters, exclamation and question marks, and emojis was calculated for both datasets across both classes (toxic vs non-toxic, hate speech vs non-hate speech). These text elements were then averaged, and comparisons were made between harmful and non-harmful language. The results are presented in Table 8.8. It was observed that, in these two datasets, sentences containing harmful language were not longer but did include more exclamation and question marks, while the number of emojis remained unchanged.

8.3.2 *Distributions*

The preliminary examination involved calculating the distributions of the new information added to the datasets. The results for SRP are presented in Table 8.9 and for MHS in Table 8.10. The tables show the frequencies per class. For the SRP the *toxic* and *very toxic* classes are merged into one.

Examining the age, we observe that both datasets indicate a higher frequency of age predictions above 50 years old. However, in the MHS dataset, there is a more balanced distribution across age groups. Regarding gender in all classes, the models predict that the texts in both datasets are more frequently authored by men compared to women. Particularly for SRP, the number of texts attributed to man surpasses that of women by more than a sixfold, according to the models. This aligns with the findings by Hilte et al. (2023), that men produce more hateful comments than women. Taking into account that such data contain instances of conservatism and regressive beliefs, or aggression, it can be assumed that the

		Toxic	Unsure	Not toxic
Age	**18-24**	6.16%	6.90%	6.25%
	25-34	17.41%	17.81%	18.36%
	35-49	16.68%	19.27%	20.83%
	50-xx	59.74%	56.03%	54.56%
Gender	**Male**	84.04%	83.27%	83.04%
	Female	15.96%	16.73%	16.96%
Sentiment	**Positive**	24.79%	28.72%	31.62%
	Negative	75.21%	71.28%	68.38%
Emotionality	**Rationality**	57.10%	64.54%	69.49%
	Emotionality	42.90%	35.46%	30.51%
Emotion	**Fear**	29.62%	28.80%	27.09%
	Disgust	27.28%	24.85%	20.43%
	Joy	13.27%	14.91%	17.89%
	Sadness	9.63%	9.73%	11.46%
	Anger	7.30%	9.29%	9.86%
	Surprise	7.09%	6.37%	6.89%
	No emotion	5.80%	6.05%	6.38%
Communication Style	**Self-revealing**	65.40%	61.71%	59.76%
	Fact-oriented	16.23%	18.81%	22.19%
	Information-seeking	13.68%	15.56%	14.48%
	Action-seeking	4.68%	15.56%	3.57%

Table 8.9: Distributions for Symanto psychoanalytics for *Jigsaw Specialized Rater Pool Dataset (SRP)*

models perceive these instances as more often written by older people, since they can appear more prone to prejudice (Álvarez Castillo et al., 2014). This could be the case also for the annotators, as they do not comprise only one age group, and that is why finding correlations between such variables (i.e., demographics and annotations) is important. It is possible that the models are biased but it must be noted that these biases derive from the early stages, namely the annotation.

In the case of communication style, self-revealing is the most prevalent com-

		Hate	Neutral	Counter
Age	**18-24**	24.68%	25.00%	25.33%
	25-34	29.32%	28.27%	27.93%
	35-49	7.13%	11.83%	13.07%
	50-xx	59.74%	34.90%	38.86%
Gender	**Male**	69.97%	65.58%	61.28%
	Female	30.03%	34.42%	38.72%
Sentiment	**Positive**	38.13%	39.65%	44.25%
	Negative	61.87%	60.35%	55.75%
Emotionality	**Rationality**	15.33%	20.54%	32.09%
	Emotionality	84.67%	79.46%	67.91%
Emotion	**Fear**	19.91%	23.03%	28.59%
	Disgust	14.70%	16.66%	16.40%
	Joy	5.11%	10.04%	17.08%
	Sadness	18.07%	14.86%	12.64%
	Anger	24.63%	19.64%	11.73%
	Surprise	15.91%	12.87%	10.58%
	No emotion	1.67%	2.90%	2.97%
Communication Style	**Self-revealing**	76.35%	74.35%	69.89%
	Fact-oriented	3.65%	6.71%	14.11%
	Information-seeking	6.52%	7.32%	7.12%
	Action-seeking	13.48%	11.63%	8.88%

Table 8.10: Distributions for Symanto psychoanalytics for *Measuring Hate Speech Dataset (MHS)*

munication style in both datasets with the rest of the styles accounting for less than 22% of the instances each in all classes. Such communication style is expected, considering hate speech and toxic language function as a way to express personal beliefs and other similar discourses are materialized through anecdotal narratives and experiences. For SRP, the majority of the texts (circa 57-69%) are considered rational by the models. The opposite occurs in MHS, where 67-84% out of the total instances, depending on the class, are marked as emotional. Concerning senti-

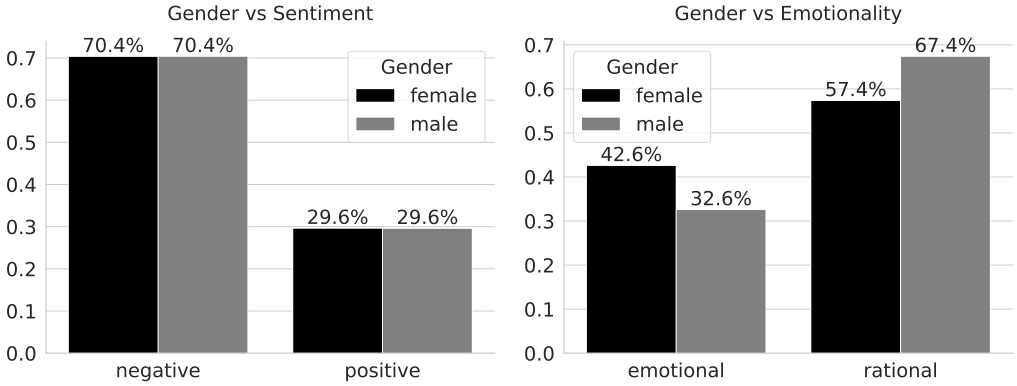


Figure 8.3: Comparison of the distributions of sentiment and emotionality in terms of gender in *Jigsaw Specialized Rater Pool Dataset (SRP)*

ment, both cases show a higher prevalence of negative texts compared to positive ones. Regarding emotions, in SRP, the most frequently predicted emotion is fear, with more than 27% of the total instances in both cases. Disgust comes second, followed by joy. In MHS, the emotion distribution is different depending on the class. For example, when no hate speech is identified, the prevalent emotion is fear, followed by joy and disgust. When we are looking at the neutral or ambiguous hate speech class, fear is again the most prevalent emotion, followed by anger and disgust. Finally, when hate speech is identified, anger is the most prevalent emotion, followed by fear, and sadness.

In general both datasets are characterized by negative sentiments and emotions rather than positive ones. One of the reasons could be the selection method of the instances during the creation of the original datasets that target specific topics and keywords (Goyal et al., 2022; Kennedy et al., 2020), resulting in instances that might not be hate speech or toxic but still contain negative nuances.

Some variables are also compared in terms of gender. In Figures 8.3 and 8.4 are presented the percentages of sentiment and emotionality in terms of gender. In SRP, we see that texts that are possibly written by both female and male authors according to the models, are associated with negative sentiment by approximately 70%, while positive sentiment is at 29.6%. In MHS, negative sentiment is more frequently associated with male authors while positive sentiment with female authors. With regard to the emotionality of the text, emotional texts are attributed more frequently to female authors compared to male ones that are attributed more rational texts. In both datasets, female authors are associated with more emotional

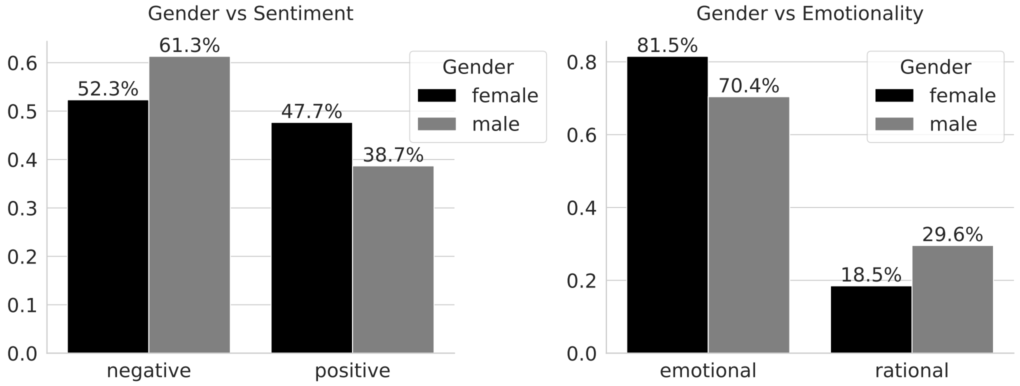


Figure 8.4: Comparison of the distributions of sentiment and emotionality in terms of gender in *Jigsaw Specialized Rater Pool Dataset (MHS)*

texts by a difference of 0.1 compared to male authors. The fact that emotionality and positive sentiment are more often attributed to female authors can be due to the fact that models are trained on data and therefore can carry biases, while it is also true from previous research that language between males and females differs and that can be picked up by the models (Lee and Kim, 2021).

8.3.3 *Regression Analysis*

In this section, the results for the coefficients of the regression analysis are presented. The estimate, the estimate error, and the credible intervals are reported. The population-level effects are then displayed separately for each factor examined in the analysis, expressed in terms of standard deviations and correlations. The first column presents the different variable groups, the second column provides the mean values (estimate), and the third column includes the standard deviation (estimate error). A goodness-of-fit test was also conducted by calculating the Rhat value (Gelman and Rubin, 1992), which indicates whether a computer simulation is performing well. For the model to be considered a good fit, the Rhat value must be lower than 1.1, which was achieved for all variables. Additionally, it was observed that lower estimates correspond to a higher probability of toxicity.

The tables are also grouped in accordance with the research questions. At this point, it needs to be noted that the lower the estimate value, the most probable is the correlation between the variable and the toxicity of the text. The significance of the results is calculated by subtracting the estimate value from the estimate error.

Emotionality and Sentiment	**Estimate**		**95% CI**	
	SRP	MHS	SRP	MHS
Rationality	$0.37_{\pm 0.01}$	$0.27_{\pm 0.01}$	[0.34, 0.40]	[0.25, 0.29]
Positive sentiment	$0.24_{\pm 0.01}$	$0.03_{\pm 0.01}$	[0.21, 0.26]	[0.02, 0.05]

Table 8.11: Coefficients for text sentiment and emotionality in SRP and MHS.

Age and Gender	**Estimate**		**95% CI**	
	SRP	MHS	SRP	MHS
Age	$-0.03_{\pm 0.02}$	$0.13_{\pm 0.01}$	[-0.07, -0.00]	[0.11, 0.14]
Men	$-0.09_{\pm 0.02}$	$-0.15_{\pm 0.01}$	[-0.12,-0.06]	[-0.16, -0.13]

Table 8.12: Coefficients for the author's age and gender in SRP and MHS.

The effect is considered significant only if the resulting absolute value is greater than the original estimate error.

The estimated coefficient for the rationality/emotionality of the text both in the SRP and MHS datasets (Table 8.11) is positive, indicating a negative correlation with the level of toxicity and hate speech of the text. The credible intervals confirm that the estimate value is true, meaning that the estimate value falls within a range of plausible values based on the statistical analysis. The results are in line with what we saw earlier when examining the distributions, as it is expected that more irrational texts would be more prone to be annotated as toxic language or hate speech, compared to rational ones.

Similarly, the estimated coefficients for the positive sentiment of the text in both datasets indicate a negative correlation with the toxicity and hate speech of the text, however, in the MHS dataset the value is very low, thus the correlation might not be so intense. A negative correlation to positive sentiment is not surprising as it is usually negative sentiment that is associated with toxic language and hate speech due to the content of such texts, that might contain pejorative wording or describe unpleasant situations and experiences.

Age and gender coefficients are presented in Table 8.12. In the SRP dataset, the estimated coefficient for age (-0.03) indicates a positive correlation with the toxicity of the text. As age here is modeled in a linear manner, the greater the age number the higher the possibility of a text to be labeled as toxic. The credible intervals show that the effect is likely to fall within the range of -0.07 to -0.00. However, in the MHS dataset the coefficient shows a negative correlation with the hate speech

Emotions	**Estimate**		**95% CI**	
	SRP	MHS	SRP	MHS
Disgust	$-0.26_{\pm 0.02}$	$0.18_{\pm 0.01}$	[-0.30, -0.21]	[0.16, 0.20]
Fear	$-0.12_{\pm 0.02}$	$0.34_{\pm 0.01}$	[-0.16, -0.07]	[0.32, 0.37]
Joy	$0.20_{\pm 0.03}$	$0.48_{\pm 0.01}$	[0.15, 0.25]	[0.45, 0.51]
No emotion	$0.21_{\pm 0.03}$	$0.30_{\pm 0.02}$	[0.15, 0.26]	[0.27, 0.35]
Sadness	$0.05_{\pm 0.03}$	$0.22_{\pm 0.01}$	[-0.00, 0.10]	[0.20, 0.25]
Surprise	$0.10_{\pm 0.03}$	$0.23_{\pm 0.01}$	[0.04, 0.17]	[0.20, 0.25]

Table 8.13: Coefficients for the emotion effects in SRP and MHS.

of the text, with an estimate of -0.13 and the credible intervals indicating that the true value falls within the range of -0.14 to -0.11. These contrasting results could be originating from the social media platforms that the datasets were extracted from. SRP was sourced from a news comments platform (Goyal et al., 2022; Borkan et al., 2019a) whereas MHS includes posts from YouTube, Twitter, and Reddit (Kennedy et al., 2020). News platforms are mainly used by elder people, who tend to produce more conservative posts that are less tolerant towards minority groups (Cornelis et al., 2009). The SRP comments might naturally be inclined to contain more toxic language, which could have impacted annotators. In addition, replicating the same experiments with two different experiments and yielding contrasting results does not necessarily mean that there is no correlation. On the contrary, we can consider it an absence of evidence. With regard to gender in the SRP dataset, the estimate for men is -0.09 indicating a positive correlation with the toxicity of the text. The credible interval indicates that the true effect size is likely to fall within the range of -0.12 to -0.06. Similarly, the estimate for the MHS dataset shows a positive correlation (0.15) with the hate speech of the text, with narrow credible intervals indicating that the true effect size lies within 0.13 and 0.16. Women use more cognitive, social and hedge words (e.g., “I think”), which conveys a linguistic style that is people-oriented, empathetic and referring to multiple perspectives. In comparison, men’s linguistic style is reported to contain more big words and swear words, and to be object- and topic-specific (Pennebaker, 2011). Thus the annotators can be more sensitive to hate speech or toxic language used by men.

Regarding the emotions (Table 8.13) that are present in the text for the SRP dataset, disgust and fear present a positive correlation with the toxicity of the text with the estimated effects being -0.26 and -0.12 respectively, values that are also included in the credible intervals. Joy, sadness, surprise, and no emotion all show negative correlation with the estimated effect for joy 0.20, for sadness 0.05, for surprise 0.10, and for no emotion 0.21. The confidence intervals for each respective emotion are 0.15 to 0.25 for joy, -0.00 to 0.10 for sadness, 0.04 to 0.17 for surprise, and finally 0.15 to 0.26 for no emotion. The MHS dataset, on the other hand, presents all negative correlation with the hate speech of the text. The coefficient for disgust is -0.26, with a credible interval of -0.20 and -0.16. The coefficient for fear is 0.34 with the credible interval 0.32 to 0.37. For joy the coefficient is 0.48 and the credible interval shows that the true effect size falls within 0.45 and 0.51. Sadness and surprise present coefficient estimates of 0.22 and 0.23, respectively, with credible intervals 0.20 to 0.25 for both variables. Finally, no emotion presents a coefficient estimate of 0.30 and a credible interval of 0.27 to 0.35. We can connect these findings with the fact that negative emotions are more likely to induce negative counter-speech. Disgust is reflected as *moral disgust* in the modern society serving for rejection of out-group ideas and values. It is explicitly directed towards groups that are considered as social norm violators. Hence, in this study, disadvantaged group members may have been triggered by being perceived as moral threat, and label disgust-related hate speech as more toxic (Ekman and Cordaro, 2011; Tybur et al., 2009). As for fear, the results for positive correlation with toxicity could be related to annotators' perception about being feared as a community. Being stigmatized as fearful and dangerous has crucial mental health impacts on minority communities (Misra et al., 2021). For annotators, this may have played a role in being selective about text that highlights fear as hate speech. The co-occurrence of disgust and fear as the two most correlated emotions is a noteworthy finding as well. These two emotions have been reported as potentially co-occurring together since they are avoidance emotions which trigger withdrawal from the stimuli, i.e. minority groups in this study (Harmon-Jones et al., 2011). This supports the idea that annotators could be impacted by emotions that stigmatize them as *groups to be avoided*, and inclined to label such text as more toxic.

For the communication style (Table 8.14) in the SRP dataset, action seeking, information seeking, and self-revealing communication styles all show positive correlations with the toxicity of the text. The estimated effects are -0.17, -0.08, and -0.09, respectively. The credible intervals show that the true effect size for action seeking is likely to fall within the range of -0.20 to -0.23, for information seeking -0.11 to -0.05, and for self-revealing -0.12 to -0.06. The fact oriented communication style presented negative correlation with the estimated effect being over the reference (0.05) and the credible intervals indicating that the true effect size lies within 0.02 and 0.08. For the MHS dataset there are only two communication

Table 8.14: Coefficients for the communication style in SRP and MHS.

Communication style	**Estimate**		**95% CI**	
	SRP	MHS	SRP	MHS
Action seeking	$-0.17_{\pm 0.02}$	$-0.11_{\pm 0.01}$	[-0.20, -0.13]	[-0.13, -0.09]
Information seeking	$-0.08_{\pm 0.01}$	$0.11_{\pm 0.01}$	[-0.11, -0.05]	[0.09, 0.13]
Self-revealing	$-0.09_{\pm 0.01}$	$-0.03_{\pm 0.01}$	[-0.12, -0.06]	[-0.05, -0.01]
Fact oriented	$0.05_{\pm 0.01}$	$0.26_{\pm 0.01}$	[0.02, 0.08]	[0.23, 0.29]

styles that are correlated to the hate of the text, i.e. action seeking and self-revealing, with the estimate being 0.11 and 0.03, respectively. The credible intervals show that the true effect for the former falls within 0.09 and 0.13, while for the latter within 0.01 and 0.05. Information seeking and fact-oriented report negative correlation. The estimate for information seeking is -0.11 with the credible intervals showing that the true effect falls within -0.13 and -0.09 while the fact-oriented communication style presents an estimate of 0.05 and credible intervals of -0.29 to -0.23. We saw in the results that in the SRP dateset, only fact oriented communication style is negatively correlated with toxicity, however slightly. All the rest are positively correlated. Fact oriented communication style was also negatively correlated in the MHS dataset. Information-seeking is also negatively correlated. These results confirm the initial hunch which expected self-revealing information to be more prone to be labeled as toxic or hate speech. Self-revealing information is about a person's life trajectory and very central to who they are including values and belief systems (Hirsh and Peterson, 2009). In the annotation task, text disclosing personal and self-relevant stories about hate speech may have drawn a picture of prejudicial and toxic traits and identities about the authors, which could have inclined annotators to label self-revealing comments as more toxic. Confirming the socio-psychological presupposition, in both datasets self-revealing texts are positively correlated with the toxicity and hate speech of the annotated texts. Hate speech is often enriched with personal stories, suggesting action and information-seeking communication styles are potentially more related to minorities, thus could have evoked more toxicity during labeling.

8.4 Discussion

This chapter has examined the correlation between hate speech and toxic labeling with two elements that could potentially result to bias: inferred demographics (age and gender) of the authors of the text, and different psycholinguistic aspects of the text (sentiment, emotion, emotionality, communication style). Most of the elements presented moderate correlation with toxicity and hate speech labeling, expanding research from previous studies that only examine annotator bias to research that takes into account the text itself as it might influence the annotator's perception of harmful language. The discussion section will use the analysis to address and answer the research questions.

RQ 1 How are psycholinguistic aspects of the text (emotionality, sentiment, emotion, communication style) associated with labeling for hate speech or toxicity purposes?

As we saw in the results, the more emotional, the higher the likelihood of the text to be labeled as toxic. This is potentially due to the fact that emotionally charged texts might trigger high levels of emotional processing (Holmes et al., 2007). In addition, the wording used in emotionally charged texts is very different from the wording used in more rational texts. Profanity, certain syntax, and other linguistic features are more likely to appear in emotional texts, facilitating the task of toxicity labeling (KhosraviNik and Esposito, 2018). Furthermore, toxicity and hate speech are phenomena that are constructed on stereotypes based on irrational assumptions and generalizations (Beeghly, 2021). Consequently, it is expected that rational language is harder to be associated with such phenomena.

With regard to sentiment, positive sentiment and toxicity or hate speech are negatively correlated in both datasets, according to the results. This is expected as toxic language and hate speech are usually characterized by negatively charged language, such as the denial of fundamental human rights accompanied by the promotion of violent and aggressive behavior, as well as the use of vulgarisms which seem to indicate hate speech (Papcunová et al., 2023). This holds true for the emotions that the texts convey in general since the results indicated that fear and disgust are correlated with toxicity and hate speech. However, annotators must be aware that a text that conveys negative sentiment or emotion is not necessarily toxic or hate speech.

The fact that communication style is also correlated with the toxicity and hate speech labeling to a certain degree, could initiate a discussion about the training of the annotators and how, maybe by being aware of such subtle differences in linguistic choices, they can avoid inducing unfair biases into their annotations.

RQ 2 Do certain demographic aspects of the author of the text (such as age and gender) affect the annotators' perception when labeling for hate speech or toxicity?

As far as age is concerned, the results from the two datasets are contrasting. For the SRP dataset, age has a positive correlation with toxicity, and given also that the linear scale is used, the greater the inferred age of the author of the text, the higher the probability of a text to be perceived as toxic by the annotators. In MHS, age shows negative correlation with the hate speech of the text. Previous research has shown that as people get older, they use more positive and less negative wording (Pennebaker and Stone, 2003). The present findings, however, contrast these studies as the numbers suggest that the older the author of the text the more likely it is to be associated with toxicity or hate speech by the annotators. Given that the study by Pennebaker and Stone (2003) was conducted in the early 2000, before the Internet started to become accessible to everyone, and before the existence of massive social media platforms like Facebook and Twitter, the state of affairs has changed calling for more longitudinal approaches. Furthermore, the Internet provides a certain degree of anonymity that gives room for the use of hateful language. Taking into account that it is more likely for older people to engage into conservatism and hateful movements, in conjunction with the anonymity of the Internet, then it is expected that this also would be the perception of the annotators, justifying the results from sociological perspective.

There were also differences in how gender is correlated with the toxicity and hate speech of the text. Considering the fact that up to a certain degree the two genders use different linguistic features, for example women tend to be more polite while men more assertive due to social conventions and expectations, it is expected that annotators perceive these differences and do not label in an unbiased manner, meaning that they might label an instance as toxic if they suspect that it is authored by a man much easier than if they suspect the author is a woman.

Overall, all the text aspects that were examined in this study might weigh into the decision-making of the annotators, particularly in borderline cases. Text aspects can appear in both toxic and non-toxic texts, but the way annotators perceive and interpret them might affect the classification.

8.5 Conclusion

This study is the first to analyze text aspects —such as inferred demographic information, emotions, and communication style of the author— to examine their effect on harmful language annotation, specifically toxicity and hate speech. All these aspects seem to be related to how annotators perceive and label toxicity and hate speech, something that could potentially introduce bias that can be harmful for corpora creation and AI model building. The study indicates that demographic text features, such as age and gender of the author, are correlated with annotators' perceptions and could be linked to differences in the annotation process, with age showing a particularly strong association in the results. Most Ekman emotions seem to be negatively correlated with toxicity, apart from fear and disgust, as well as negative sentiment indicating that annotators might be more prone to labeling a text as hate speech or toxic if the sentiment or emotion the text conveys is negative.

Regarding communication style, fact oriented presented negative correlation in both datasets, while most of the rest of communication styles were positively correlated. This analysis has provided us with a few insights on potential biases of the perception of hate speech, paving the way for more research that will lead to more thorough data collection and that will take into account more variables that concern the text and not only the demographics of the annotator.

Chapter 9

Conclusions

This thesis has explored automatic hate speech detection, addressing key challenges such as the selection and application of definitions, bias mitigation, generalization, and cross-lingual dataset creation. The primary contributions of this work include the development of datasets—either as re-annotated versions of existing data or entirely new collections—and providing valuable insights into the appropriate framework selection for specific tasks in hate speech detection.

To proceed to the experimental part of this thesis, I first analyzed the current state of automatic hate speech detection. Chapter 3 highlighted that defining hate speech and related concepts, such as toxicity and abusiveness, remains an open issue, even for legislative authorities that attempt to define hate speech and provide laws. The appropriate selection or creation of definitions is heavily task-dependent and must align with the annotation guidelines established by researchers.

This lead to an examination of current annotation schemes which are another subject of ongoing debate. Given the inherently subjective nature of hate speech detection, several questions arise: Should annotators always be provided with explicit definitions? Should they possess expert knowledge in the field? Furthermore, annotation results can be influenced by personal biases and contextual factors, such as the style of the text or linguistic cues that suggest the author's demographics (e.g., gender) or attitudes (e.g., communication style or emotional tone).

Finally, Chapter 3 also underscored that current models face significant challenges in tackling issues like generalization, multilinguality, and cultural awareness. Existing approaches still fall short in addressing these challenges, leaving room for further improvements in the field. This thesis has attempted to address aspects of these challenges.

First, focusing on the issue of defining hate speech and related concepts—and how these definitions influence annotation and generalization—I presented a re-annotation experiment in Chapter 4. This experiment evaluates the robustness of existing datasets by comparing their original labels with re-annotated ones in terms of annotator agreement. Additionally, I conducted BERT-based model experiments, evaluating the model's performance on both the original labels and the re-annotations. The human annotation shows that, although in most cases the annotations were inconsistent, Toxicity and HOTA (any of the following: Hateful, Offensive, Toxic, Abusive language) appear to lead to higher and more consistent inter-annotator agreement. The experimental model, on the other hand, showed that, assessing on data from the same source as the training set, when using the original ground truth, can lead to higher accuracy compared to assessing data from a different source, confirming previous studies. Yet, this cannot be used as a rule of thumb since testing on the re-annotations showed that the performance can drop when testing on the data from the same source as the training set and it can increase when testing on previously completely unseen data. Future work could examine more datasets and also leverage the power of LLMs for the model experiments.

As the interpretation of the definitions used in Chapter 4 proved highly subjective, I investigated whether the same applies to hate speech laws. In Chapter 5, I presented a newly judged sample dataset on prosecutable hate speech. Using this labeled data, I conducted experiments to predict whether a given hate speech instance could be considered prosecutable. Additionally, Qwen2 was employed to generate silver data, which was evaluated on the re-annotated dataset. The results of the re-annotation demonstrated that, even with explicit reference to legislation, determining whether an instance qualifies as prosecutable hate speech remains a matter of subjective interpretation. Furthermore, the models used in the experiments tended to be more conservative, predicting fewer instances as prosecutable compared to non-prosecutable. Interestingly, this tendency aligns with human annotations, where both experts and models demonstrated a bias toward the "Not Prosecutable" class. Our error analysis further revealed that multiclass PLMs trained with leave-one-out cross-validation are more effective for prosecutable hate speech detection than LLMs. Finally, LLM-generated data did not contribute to improved performance for PLMs. In future work, the re-annotated dataset could be expanded with additional annotations based on hate speech laws from different countries, providing broader overview of how experts perceive prosecutable hate speech and fostering the development of more robust and more culturally aware models.

As part of future work is exploring the impact of different definitions of hate speech, Chapter 6 introduced `HateDefCon`, a comprehensive dataset dedicated to hate speech definitions. HateDefCon provides detailed annotations for components such as the target, intent/purpose, and act/means associated with hate speech instances, enabling a comparative analysis of hate speech definitions. The analysis of HateDefCon revealed significant gaps in cultural diversity within existing definitions. Most definitions were either in English, translated from English, or ambiguous regarding their cultural origins, as platforms often tied definitions to a language rather than a specific culture. Legislative sources were the only ones explicitly linked to particular countries or cultural contexts. The components of hate speech definitions showed significant variation, even in contexts, such as with Levantine Arabic, where one language spans multiple cultural backgrounds. The model experiments demonstrated that definitions significantly influence the performance of LLMs. In many cases, more comprehensive definitions led to better performance. This finding underscores two critical considerations for hate speech detection research:

- Definitions incorporated into annotation guidelines or prompts must be tailored to the specific task, with clear indications of whether a general or comprehensive definition is required. Furthermore, model selection should align with the definition type, as some models perform better with general definitions while others perform better with more detailed ones.

- Definitions must be aligned with the language and culture of the target dataset. This could involve referencing hate speech legislation to clarify what constitutes hate speech within a given cultural context. Where cultural specificity is not applicable, it should be explicitly stated that no specific culture is considered.

Future work could develop a community-driven knowledge base of hate speech definitions. This resource will support research across a broader range of cultures and domains, simultaneously addressing emerging challenges.

Chapter 7 addressed several challenges involved in creating a parallel hate speech corpus. Specifically, I used machine translation to translate English tweets into Greek and Italian, followed by backtranslation combined with translation metrics to evaluate the quality of the translations. I also conducted tests to assess changes in toxicity levels in the translated texts and performed a qualitative analysis on a sample of the instances used. The findings indicate that while machine translation can adequately preserve the intended meaning of sentences, it often introduces grammatical and syntactical errors, making many translations unsuitable for inclusion in a parallel corpus. Only a small number of examples successfully retained both the original meaning and toxicity levels without grammatical or syntactical issues, highlighting the need for further refinement of the pipeline. Future work can focus expanding the experiments with a broader range of languages and incorporate English as a target language to investigate whether the observed decrease in toxicity levels is consistent across other cases.

Chapter 8 of this thesis examined bias stemming from textual aspects of harmful language. This study analyzed various text features—such as inferred demographic information, emotions, and the author's communication style—to investigate their impact on the annotation of harmful language, specifically toxicity and hate speech. The findings suggest that these aspects influence how annotators perceive and label such content, potentially introducing biases that could adversely affect both corpus creation and AI model development. The study revealed that demographic features inferred from the text, such as the author's age and gender, were correlated with annotators' perceptions. Age, in particular, exhibited a strong association with annotation differences. Emotional content also played a significant role: most of Ekman's emotions were negatively correlated with toxicity, except for fear and disgust, which, along with negative sentiment, were positively associated. This suggests that annotators may be more likely to label a text as toxic or hate speech if it conveys negative emotions or sentiment. Regarding communication style, fact-oriented texts showed a negative correlation with toxicity labels in both datasets, whereas most other communication styles were positively correlated. Future research should aim to improve data collection processes by considering additional textual variables, alongside annotator demographics, to mitigate bias and encourage the development of more accurate models.

In summary, the practical outcomes of this thesis include two re-annotated datasets: one annotated with the definition of hate speech and other related definitions by crowd-source workers, and one annotated for prosecutable hate speech by legal experts. In addition, the first dataset of hate speech definitions accompanied by an annotation framework designed for cross-domain and cross-cultural comparisons is introduced, as well as a parallel multilingual hate speech dataset along with its creation pipeline. The thesis also provides an analysis of textual aspects, focusing on how annotators perceive them and how these perceptions influence the annotation process. It is hoped that these contributions will inspire further research. The provided resources are intended to be a foundation for future efforts, encouraging the community to refine and experiment with them.

Bibliography

AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.alw-1.21.

Manuel Alcántara-Plá. Understanding emotions in hate speech: A methodology for discourse analysis. *Discourse & Society*, 35(4):417–433, 2024. doi: 10.1177/09579265231222013. URL https://doi.org/10.1177/09579265231222013.

Natalie Alkiviadou, Jacob Mchangama, and Raghav Mendiratta. *Global Handbook on Hate Speech Laws*. Justitia, Denmark, 2020. © Justitia and the authors, 2020.

Pedro Alonso, Rajkumar Saini, and György Kovacs. TheNorth at SemEval-2020 task 12: Hate speech detection using RoBERTa. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2197–2202, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.292. URL https://aclanthology.org/2020.semeval-1.292.

Sai Saket Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*, 2020.

José Luis Álvarez Castillo, Alfredo Jiménez Equizábal, Carmen Palmero Cámara, and Hugo González González. The fight against prejudice in older adults:

Perspective taking effectiveness. *Revista Latinoamericana de Psicología*, 46 (1-3):137–147, 2014. ISSN 0120-0534. doi: 10.1016/S0120-0534(14)70017-2.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Z. Chen, Eric Chu, J. Clark, Laurent El Shafey, Yanping Huang, Kathleen S. Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernández Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Michael Brooks, Michele Catasta, Yongzhou Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, C Crépy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, M. C. D'iaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fan Feng, Vlad Fienber, Markus Freitag, Xavier García, Sebastian Gehrmann, Lucas González, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, An Ren Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wen Hao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Mu-Li Li, Wei Li, Yaguang Li, Jun Yu Li, Hyeontaek Lim, Han Lin, Zhong-Zhong Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Oleksandr Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alexandra Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Marie Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniela Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Ke Xu, Yunhan Xu, Lin Wu Xue, Pengcheng Yin, Jiahui Yu, Qiaoling Zhang, Steven Zheng, Ce Zheng, Wei Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. Palm 2 technical report. *ArXiv*, abs/2305.10403, 2023. URL https://api.semanticscholar.org/CorpusID:258740735.

Aymé Arango, Jorge Pérez, and Barbara Poblete. Cross-lingual hate speech detection based on multilingual domain-specific word embeddings. *CoRR*, abs/2104.14728, 2021. URL https://arxiv.org/abs/2104.14728.

Carlos Arcila-Calderón, Javier J. Amores, Patricia Sánchez-Holgado, Lazaros Vrysis, Nikolaos Vryzas, and Martín Oller Alonso. How to detect online hate towards migrants and refugees? Developing and evaluating a classifier of racist and xenophobic hate speech using shallow and deep learning. *Sustainability*, 14(20), 2022. ISSN 2071-1050. doi: 10.3390/su142013094. URL https://www.mdpi.com/2071-1050/14/20/13094.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.c3nlp-1.12. URL https://aclanthology.org/2023.c3nlp-1.12.

Stavros Assimakopoulos, Rebecca Vella Muskat, Lonneke van der Plas, and Albert Gatt. Annotating for hate speech: The MaNeCo corpus and some input from critical discourse analysis. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5088–5097, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.626.

Sunyam Bagga and Andrew Piper. Measuring the effects of bias in training data for literary classification. In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 74–84, Online, December 2020. International Committee on Computational Linguistics. URL https://aclanthology.org/2020.latechclfl-1.9.

Agathe Balayn, Jie Yang, Zoltan Szlavik, and Alessandro Bozzon. Automatic identification of harmful, aggressive, abusive, and offensive language on the web: A survey of technical biases informed by psychology literature. *Trans. Soc. Comput.*, 4(3), 2021. ISSN 2469-7818. doi: 10.1145/3479158.

Alberto Barrón-Cedeño, Cristina España-Bonet, Josu Boldoba, and Lluís Màrquez. A factory of comparable corpora from Wikipedia. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 3–13, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3402. URL https://aclanthology.org/W15-3402.

Matthew Barthet, Chintan Trivedi, Kosmas Pinitas, Emmanouil Xylakis, Konstantinos Makantasis, Antonios Liapis, and Georgios Yannakakis. Knowing your annotator: Rapidly testing the reliability of affect annotation, 08 2023.

Angelo Basile, Guillermo Pérez-Torró, and Marc Franco-Salvador. Probabilistic ensembles of zero-and few-shot learning models for emotion classification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 128–137, Online, 2021.

Valerio Basile. It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *DP@AI*IA*, 2020. URL https://api.semanticscholar.org/CorpusID:229344921.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2007. URL https://aclanthology.org/S19-2007.

Christin Beck, Hannah Booth, Mennatallah El-Assady, and Miriam Butt. Representation problems in linguistic annotations: Ambiguity, variation, uncertainty, error and bias. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 60–73, Barcelona, Spain, 2020. Association for Computational Linguistics.

Djamila Romaissa Beddiar, Md Saroar Jahan, and Mourad Oussalah. Data expansion using back translation and paraphrasing for hate speech detection. *Online Social Networks and Media*, 24:100153, 2021. ISSN 2468-6964.

Erin Beeghly. What's wrong with stereotypes? The falsity hypothesis. *Social Theory and Practice*, 47(1):33–61, 2021.

Lisa Beinborn and Rochelle Choenni. Semantic Drift in Multilingual Representations. *Computational Linguistics*, 46(3):571–603, 11 2020. ISSN 0891-2017. doi: 10.1162/coli_a_00382. URL https://doi.org/10.1162/coli_a_00382.

Jonah Berger and Katherine L Milkman. Emotion and virality: what makes online content go viral? *NIM Marketing Intelligence Review*, 5(1):18–23, 2013.

Nicola Bertoldi, Davide Caroselli, and Marcello Federico. The ModernMT project. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, page 365, Alicante, Spain, May 2018. URL https://aclanthology.org/2018.eamt-main.46.

Irina Bigoulaeva, Viktor Hangya, and Alexander Fraser. Cross-lingual transfer learning for hate speech detection. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 15–25, Kyiv, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.ltedi-1.3.

Irina Bigoulaeva, Viktor Hangya, Iryna Gurevych, and Alexander Fraser. Addressing the challenges of cross-lingual hate speech detection. *CoRR*, abs/2201.05922, 2022. URL https://arxiv.org/abs/2201.05922.

Irina Bigoulaeva, Viktor Hangya, Iryna Gurevych, and Alexander Fraser. Label modification and bootstrapping for zero-shot cross-lingual hate speech detection. *Language Resources and Evaluation*, 57:1515–1546, 2023. doi: 10.1007/s10579-023-09637-4. URL https://doi.org/10.1007/s10579-023-09637-4.

Victoria Bobicev and Marina Sokolova. Inter-annotator agreement in sentiment analysis: Machine learning perspective. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 97–102, Varna, Bulgaria, 2017. INCOMA Ltd. doi: 10.26615/978-954-452-049-6_015.

Robert J Boeckmann and Jeffrey Liew. Hate speech: Asian american students' justice judgments and psychological responses. *Journal of Social Issues*, 58(2): 363–381, 2002.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019a.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. *CoRR*, abs/1903.04561, 2019b.

Tom Bourgeade, Patricia Chiril, Farah Benamara, and Véronique Moriceau. What did you learn to hate? a topic-oriented analysis of generalization in hate speech detection. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of* *the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3495–3508, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.254. URL https://aclanthology.org/2023.eacl-main.254.

Daniel Braun. I beg to differ: How disagreement is handled in the annotation of legal machine learning data sets. *Artif Intell Law*, 2023. doi: 10.1007/s10506-023-09369-4.

Alexander Brown. What is hate speech? part 2: Family resemblances. *Law and Philosophy*, 36:561–613, 2017. doi: 10.1007/s10982-017-9300-x. URL https://doi.org/10.1007/s10982-017-9300-x.

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. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Paul-Christian Bürkner. brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1):1–28, 2017. doi: 10.18637/jss.v080.i01.

Federico Cabitza, Andrea Campagner, and Valerio Basile. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868, Jun. 2023. doi: 10.1609/aaai.v37i6.25840. URL https://ojs.aaai.org/index.php/AAAI/article/view/25840.

Dana R Carney, C Randall Colvin, and Judith A Hall. A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality*, 41(5):1054–1072, 2007.

Marco Casavantes, Mario Ezra Aragón, Luis Carlos González, and Manuel Montes. Leveraging posts' and authors' metadata to spot several forms of abusive comments in twitter. *Journal of Intelligent Information Systems*, 61:519–539, 2023. doi: 10.1007/s10844-023-00779-z. URL `https://doi.org/10.1007/s10844-023-00779-z`.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.woah-1.3.

Camilla Casula, Elisa Leonardelli, and Sara Tonelli. Don‘t augment, rewrite? assessing abusive language detection with synthetic data. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11240–11247, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.669. URL `https://aclanthology.org/2024.findings-acl.669/`.

Claudia Casula and Sara Tonelli. Hate speech detection with machine-translated data: The role of annotation scheme, class imbalance and undersampling. In Felice Dell'Orletta, Johanna Monti, and Fabio Tamburini, editors, *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020*, pages 1–. Accademia University Press, 2020. doi: 10.4000/books.aaccademia.8345. URL `https://doi.org/10.4000/books.aaccademia.8345`.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.261.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3), March 2024. ISSN 2157-6904. doi: 10.1145/3641289. URL `https://doi.org/10.1145/3641289`.

Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Measuring #gamergate: A tale of hate, sexism, and bullying. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 1285–1290, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee. ISBN 9781450349147. doi: 10.1145/3041021.3053890.

Hung Chau, Saeid Balaneshin, Kai Liu, and Ondrej Linda. Understanding the tradeoff between cost and quality of expert annotations for keyphrase extraction. In Stefanie Dipper and Amir Zeldes, editors, *Proceedings of the 14th Linguistic Annotation Workshop*, pages 74–86, Barcelona, Spain, December 2020. Association for Computational Linguistics. URL `https://aclanthology.org/2020.law-1.7`.

Lu Cheng, Ruocheng Guo, and Huan Liu. Robust cyberbullying detection with causal interpretation. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, pages 169–175, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366755. doi: 10.1145/3308560.3316503.

Naganna Chetty and Sreejith Alathur. Hate speech review in the context of online social networks. *Aggression and Violent Behavior*, 40:108–118, 2018. ISSN 1359-1789. doi: https://doi.org/10.1016/j.avb.2018.05.003.

Mara Chinea-Rios, Thomas Müller, Gretel Liz De la Peña Sarracén, Francisco Rangel, and Marc Franco-Salvador. Zero and few-shot learning for author profiling. In *Natural Language Processing and Information Systems: 27th International Conference on Applications of Natural Language to Information Systems, NLDB 2022*, pages 333–344, Valencia, Spain, 2022. Springer.

Patricia Chiril, Endang Wahyu Pamungkas, Farah Benamara, et al. Emotionally informed hate speech detection: A multi-target perspective. *Cognitive Computation*, 14:322–352, 2022. doi: 10.1007/s12559-021-09862-5.

Anna Chmiel, Pawel Sobkowicz, Julian Sienkiewicz, Georgios Paltoglou, Kevan Buckley, Mike Thelwall, and Janusz A Hołyst. Negative emotions boost user activity at bbc forum. *Physica A: statistical mechanics and its applications*, 390 (16):2936–2944, 2011.

Ilse Cornelis, Alain Van Hiel, Arne Roets, and Malgorzata Kossowska. Age differences in conservatism: Evidence on the mediating effects of personality and cognitive style. *Journal of personality*, 77(1):51–88, 2009.

Catherine A. Cottrell and Steven L. Neuberg. Different emotional reactions to different groups: a sociofunctional threat-based approach to" prejudice". *Journal of personality and social psychology*, 88(5):770, 2005.

Council of Europe. Recommendation no. r (97) 20 of the committee of ministers to member states on "hate speech". Retrieved from https://rm.coe.int/1680505d5b, 1997. URL `https://rm.coe.int/1680505d5b`.

Gloria Cowan and Cyndi Hodge. Judgments of hate speech: The effects of target group, publicness, and behavioral responses of the target. *Journal of Applied Social Psychology*, 26:355–374, 1996.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319, 2023. doi: 10.1162/tacl_a_00550. URL `https://aclanthology.org/2023.tacl-1.18`.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language, 2017a.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. *CoRR*, abs/1703.04009, 2017b.

Reinout E de Vries, Angelique Bakker-Pieper, Femke E Konings, and Barbara Schouten. The communication styles inventory (csi) a six-dimensional behavioral model of communication styles and its relation with personality. *Communication Research*, 40(4):506–532, 2013.

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

James Donato. Dworkin and subjectivity in legal interpretation. *Stanford Law Review*, 40(6):1517–1541, 1988.

John Duckitt. A dual-process cognitive-motivational theory of ideology and prejudice. *Advances in Experimental Social Psychology*, 33:41–113, 2001.

Paul Ekman and Daniel Cordaro. What is meant by calling emotions basic. *Emotion review*, 3(4):364–370, 2011.

Paul Ekman and Wallace V Friesen. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *semiotica*, 1(1):49–98, 1969.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. Hate lingo: A target-based linguistic analysis of hate speech in social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1), 2018a. doi: 10.1609/icwsm.v12i1.15041.

Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. Peer to peer hate: Hate speech instigators and their targets. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1), 2018b. doi: 10.1609/icwsm.v12i1.15038.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. Latent hatred: A benchmark for understanding implicit hate speech. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.29. URL `https://aclanthology.org/2021.emnlp-main.29`.

EU Commission. Code of conduct on countering illegal hate speech online. Retrieved from https://ec.europa.eu, 2016. URL `https://ec.europa.eu`.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *EVALITA@CLiC-it*, 2018.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. SemEval-2022 task 5: Multimedia automatic misogyny identification. In Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth

Singh, and Shyam Ratan, editors, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.semeval-1.74. URL https://aclanthology.org/2022.semeval-1.74.

Audrey Fino. Defining hate speech: A seemingly elusive task. *Journal of International Criminal Justice*, 18(1):31–57, 06 2020. ISSN 1478-1387. doi: 10.1093/jicj/mqaa023. URL https://doi.org/10.1093/jicj/mqaa023.

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in Slovene. In *Proceedings of the First Workshop on Abusive Language Online*, pages 46–51, Vancouver, BC, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3007. URL https://aclanthology.org/W17-3007.

Lucie Flekova, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel Preoţiuc-Pietro. Analyzing biases in human perception of user age and gender from text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 843–854, 2016.

Caterina Flick. The legal framework on hate speech and the internet good practices to prevent and counter the spread of illegal hate speech online. In *Language, Gender and Hate Speech A Multidisciplinary Approach*. Fondazione Università Ca' Foscari, dec 2020. doi: 10.30687/978-88-6969-478-3/011. URL https://doi.org/10.30687%2F978-88-6969-478-3%2F011.

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

Paula Fortuna, Juan Soler, and Leo Wanner. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.838.

Paula Fortuna, Juan Soler-Company, and Leo Wanner. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524, 2021. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2021.102524. URL https://www.sciencedirect.com/science/article/pii/S0306457321000339.

Paula Fortuna, Monica Dominguez, Leo Wanner, and Zeerak Talat. Directions for NLP practices applied to online hate speech detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11794–11805, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12, 2018.

Simona Frenda. The role of sarcasm in hate speech.a multilingual perspective. In *Proceedings of the Doctoral Symposium of the XXXIV International Conference of the Spanish Society for Natural Language Processing*, Seville, Spain, 2018.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, September 2024. doi: 10.1162/coli_a_00524. URL https://aclanthology.org/2024.cl-3.8/.

Lei Gao, Alexis Kuppersmith, and Ruihong Huang. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In Greg Kondrak and Taro Watanabe, editors, *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 774–782, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I17-1078.

Dirk Geeraerts. Componential analysis. In Keith Brown, editor, *Encyclopedia of Language & Linguistics (Second Edition)*, pages 709–712. Elsevier, Oxford, second edition edition, 2006. ISBN 978-0-08-044854-1. doi: https://doi.org/10.1016/B0-08-044854-2/01029-4. URL https://www.sciencedirect.com/science/article/pii/B0080448542010294.

Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457 – 472, 1992. doi: 10.1214/ss/1177011136.

Phyllis B. Gerstenfeld. A time to hate: Situational antecedents of intergroup bias. *Analyses of Social Issues and Public Policy*, 2(1):61–67, 2002. doi: https://doi.org/10.1111/j.1530-2415.2002.00027.x.

Ine Gevers, Ilia Markov, and Walter Daelemans. Linguistic analysis of toxic language on social media. *Computational Linguistics in the Netherlands Journal*, 12:33–48, 2022.

Dan Gillick and Yang Liu. Non-expert evaluation of summarization systems is risky. In Chris Callison-Burch and Mark Dredze, editors, *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148–151, Los Angeles, June 2010. Association for Computational Linguistics. URL https://aclanthology.org/W10-0722.

Vaishali U. Gongane, Mousami V. Munot, and Alwin D. Anuse . Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining*, 12(1):129, 2022. doi: 10.1007/s13278-022-00951-3. URL https://doi.org/10.1007/s13278-022-00951-3.

Ward H. Goodenough. Componential analysis and the study of meaning. *Language*, 32(1):195–216, 1956. ISSN 00978507, 15350665. URL http://www.jstor.org/stable/410665.

Nitesh Goyal, Ian D. Kivlichan, Rachel Rosen, and Lucy Vasserman. Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. *Proceedings of the ACM on Human-Computer Interaction*, 6:1 – 28, 2022.

Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. All you need is "love": Evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, AISec '18, page 2–12, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360043. doi: 10.1145/3270101.3270103. URL https://doi.org/10.1145/3270101.3270103.

Joshua Guberman and Libby Hemphill. Challenges in modifying existing scales for detecting harassment in individual tweets. In *Hawaii International Conference on System Sciences*, 2017.

Victoria Guillén-Nieto. *Hate Speech*. De Gruyter Mouton, Berlin, Boston, 2023. ISBN 9783110672619. doi: doi:10.1515/9783110672619. URL https://doi.org/10.1515/9783110672619.

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

Eddie Harmon-Jones, Cindy Harmon-Jones, David M Amodio, and Philip A Gable. Attitudes toward emotions. *Journal of personality and social psychology*, 101 (6):1332, 2011.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.482. URL https://aclanthology.org/2022.acl-long.482.

Mika Hietanen and Johan Eddebo. Towards a definition of hate speech—with a focus on online contexts. *Journal of Communication Inquiry*, 47(4):440–458, 2023a. doi: 10.1177/01968599221124309. URL https://doi.org/10.1177/01968599221124309.

Mika Hietanen and Johan Eddebo. Towards a definition of hate speech—with a focus on online contexts. *Journal of Communication Inquiry*, 47(4):440–458, 2023b. doi: 10.1177/01968599221124309. URL https://doi.org/10.1177/01968599221124309.

Lisa Hilte, Ilia Markov, Nikola Ljubešić, Darja Fišer, and Walter Daelemans. Who are the haters? a corpus-based demographic analysis of authors of hate speech. *Frontiers in Artificial Intelligence*, 6, 2023. ISSN 2624-8212. doi:

10.3389/frai.2023.986890. URL https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2023.986890.

Wolfram Hinzen, Edouard Machery, and Markus Werning. *The Oxford Handbook of Compositionality*. Oxford University Press, 02 2012. ISBN 9780199541072. doi: 10.1093/oxfordhb/9780199541072.001.0001. URL https://doi.org/10.1093/oxfordhb/9780199541072.001.0001.

Jacob B Hirsh and Jordan B Peterson. Personality and language use in self-narratives. *Journal of research in personality*, 43(3):524–527, 2009.

Danielle Holmes, Georg W. Alpers, Tasneem Ismailji, Catherine Classen, Talor Wales, Valerie Cheasty, Andrew Miller, and Cheryl Koopman. Cognitive and emotional processing in narratives of women abused by intimate partners. *Violence Against Women*, 13(11):1192–1205, 2007. doi: 10.1177/1077801207307801. PMID: 17951592.

Shi Yin Hong and Susan Gauch. Improving cross-domain hate speech generalizability with emotion knowledge. In Chu-Ren Huang, Yasunari Harada, Jong-Bok Kim, Si Chen, Yu-Yin Hsu, Emmanuele Chersoni, Pranav A, Winnie Huiheng Zeng, Bo Peng, Yuxi Li, and Junlin Li, editors, *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 282–292, Hong Kong, China, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.paclic-1.29.

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.67. URL https://aclanthology.org/2023.findings-acl.67.

Md Saroar Jahan and Mourad Oussalah. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232, 2023. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.126232. URL https://www.sciencedirect.com/science/article/pii/S0925231223003557.

Md Saroar Jahan, Djamila Romaissa Beddiar, Mourad Chabane Oussalah, Nabil Arhab, and Yazid Bounab. Hate and offensive language detection using bert for english subtask a. In *Fire*, 2021. URL https://api.semanticscholar.org/CorpusID:251020400.

Md Saroar Jahan, Mourad Oussalah, and Nabil Arhab. Finnish hate-speech detection on social media using CNN and FinBERT. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 876–882, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.92.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *CoRR*, abs/2310.06825, 2023. doi: 10.48550/ARXIV.2310.06825. URL https://doi.org/10.48550/arXiv.2310.06825.

Tingting Jiang, Jaimie L Gradus, and Anthony J Rosellini. Supervised machine learning: A brief primer. *Behavior Therapy*, 51(5):675–687, 2020. doi: 10.1016/j.beth.2020.05.002. URL https://doi.org/10.1016/j.beth.2020.05.002.

Jigsaw. Jigsaw toxic comment classification challenge. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge, 2019. Accessed: [8 May 2023].

Google Jigsaw. Perspective api, 2017. URL https://www.perspectiveapi.com/.

Yiping Jin, Leo Wanner, and Alexander Shvets. GPT-HateCheck: Can LLMs write better functional tests for hate speech detection? In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7867–7885, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.694.

Hwiyeol Jo, Soo-Min Kim, and Jeong Ryu. What we really want to find by sentiment analysis: The relationship between computational models and psychological state. *ArXiv*, abs/1704.03407, 2017.

Mladen Karan and Jan Šnajder. Cross-domain detection of abusive language online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5117. URL https://aclanthology.org/W18-5117.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Alyzeh Hussain, Austin Lara, Gabriel Cardenas, Adam Omary, Christina Park, Xin Wang, Clarisa Wijaya, Yong Zhang, Beth Meyerowitz, and Morteza Dehghani. Introducing the Gab Hate Corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56(1):79–108, March 2022. ISSN 1574-0218. doi: 10.1007/s10579-021-09569-x. URL https://doi.org/10.1007/s10579-021-09569-x.

Chris J. Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *CoRR*, abs/2009.10277, 2020.

Majid KhosraviNik and Eleonora Esposito. Online hate, digital discourse and critique: Exploring digitally-mediated discursive practices of gender-based hostility. *Lodz Papers in Pragmatics*, 14(1):45–68, 2018. doi: doi:10.1515/lpp-2018-0003. URL https://doi.org/10.1515/lpp-2018-0003.

Urja Khurana, Ivar Vermeulen, Eric Nalisnick, Marloes Van Noorloos, and Antske Fokkens. Hate speech criteria: A modular approach to task-specific hate speech definitions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 176–191, Seattle, Washington (Hybrid), July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.woah-1.17. URL https://aclanthology.org/2022.woah-1.17.

David B. Kronenfeld. Cognitive research methods. In Kimberly Kempf-Leonard, editor, *Encyclopedia of Social Measurement*, pages 361–374. Elsevier, New York, 2005. ISBN 978-0-12-369398-3. doi: https://doi.org/10.1016/B0-12-369398-5/00325-X. URL https://www.sciencedirect.com/science/article/pii/B012369398500325X.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. Benchmarking aggression identification in social media. In *Proceedings of the First*

Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), pages 1–11, Santa Fe, New Mexico, USA, August 2018a. Association for Computational Linguistics. URL https://aclanthology.org/W18-4401.

Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. Aggression-annotated corpus of Hindi-English code-mixed data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018b. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1226.

Ritesh Kumar, Ojaswee Bhalla, Madhu Vanthi, Shehlat Maknoon Wani, and Siddharth Singh. HarmPot: An annotation framework for evaluating offline harm potential of social media text. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8016–8034, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.706.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521 (7553):436–444, 2015.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Juho Kim, and Alice Oh. Crehate: Cross-cultural re-annotation of english hate speech dataset, 2023a.

Nayeon Lee, Chani Jung, and Alice Oh. Hate speech classifiers are culturally insensitive. In Sunipa Dev, Vinodkumar Prabhakaran, David Adelani, Dirk Hovy, and Luciana Benotti, editors, *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 35–46, Dubrovnik, Croatia, May 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.c3nlp-1.5. URL https://aclanthology.org/2023.c3nlp-1.5.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4205–4224, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

doi: 10.18653/v1/2024.naacl-long.236. URL https://aclanthology.org/2024.naacl-long.236.

Yong-Hun Lee and Ji-Hye Kim. A sentiment analysis of men's and women's speech in the BNC64. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 603–610, Shanghai, China, 2021. Association for Computational Lingustics.

Geoffrey Leech. *Semantics: The Study of Meaning*. A Penguin book. Penguin Books, 1990. ISBN 9780140134872. URL https://books.google.it/books?id=Z9m4uQAACAAJ.

Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. A new generation of perspective api: Efficient multilingual character-level transformers, 2022. URL https://arxiv.org/abs/2202.11176.

Shui-Kwan Leung and Michael Harris Bond. Interpersonal communication and personality: Self and other perspectives. *Asian Journal of Social Psychology*, 4 (1):69–86, 2001.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

Holly Lopez Long, Alexandra O'Neil, and Sandra Kübler. On the interaction between annotation quality and classifier performance in abusive language detection. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 868–875, Held Online, September 2021. INCOMA Ltd. URL https://aclanthology.org/2021.ranlp-1.99.

Floyd G. Lounsbury. A semantic analysis of the pawnee kinship usage. *Language*, 32(1):158–194, 1956. ISSN 00978507, 15350665. URL http://www.jstor.org/stable/410664.

Florian Ludwig, Klara Dolos, Torsten Zesch, and Eleanor Hobley. Improving generalization of hate speech detection systems to novel target groups via domain adaptation. In Kanika Narang, Aida Mostafazadeh Davani, Lambert Mathias, Bertie Vidgen, and Zeerak Talat, editors, *Proceedings of the Sixth Workshop on*

Online Abuse and Harms (WOAH), pages 29–39, Seattle, Washington (Hybrid), July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.woah-1.4. URL https://aclanthology.org/2022.woah-1.4.

Chu Luo, Rohan Bhambhoria, Samuel Dahan, and Xiaodan Zhu. Legally enforceable hate speech detection for public forums. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10948–10963, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.730. URL https://aclanthology.org/2023.findings-emnlp.730.

C Neil Macrae and Susanne Quadflieg. Perceiving people handbook of social psychology, 2010.

Arun S. Maiya. ktrain: A low-code library for augmented machine learning. *arXiv preprint arXiv:2004.10703*, 2020.

Thomas Mandla, Sandip Modha, Gautam Kishore Shahi, Amit Kumar Jaiswal, Durgesh Nandini, Daksh Patel, Prasenjit Majumder, and Johannes Schäfer. Overview of the hasoc track at fire 2020: Hate speech and offensive content identification in indo-european languages, 2021. URL https://arxiv.org/abs/2108.05927.

Alice E. Marwick and Ross Miller. Online harassment, defamation, and hateful speech: A primer of the legal landscape. Report 2, Fordham Center on Law and Information Policy, 2014. URL https://ssrn.com/abstract=2447904.

B. Mathew, R. Dutt, P. Goyal, and A. Mukherjee. Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*, page 173–182, New York, NY, USA, 2019.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. 2021.

Marcel Maussen and Ralph Grillo. Regulation of speech in multicultural societies: Introduction. *Journal of Ethnic and Migration Studies*, 40(2):174–193, 2014. doi: 10.1080/1369183X.2013.851470.

Gretchen McCulloch. *Because Internet: Understanding the New Rules of Language*. Penguin Books, 2019. ISBN 9781529112825. URL https://www.penguin.

co.uk/books/441467/because-internet-by-gretchen-mcculloch/9781529112825.

Jack McDevitt, Jack Levin, and Susan Bennett. Hate crime offenders: An expanded typology. *Journal of Social Issues*, 58(2):303–317, 2002. doi: https://doi.org/10.1111/1540-4560.00262.

Yelena Mejova, Amy X Zhang, Nicholas Diakopoulos, and Carlos Castillo. Controversy and sentiment in online news. *arXiv preprint arXiv:1409.8152*, 2014.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. URL https://arxiv.org/abs/1301.3781.

Changrong Min, Hongfei Lin, Ximing Li, He Zhao, Junyu Lu, Liang Yang, and Bo Xu. Finding hate speech with auxiliary emotion detection from self-training multi-label learning perspective. *Information Fusion*, 96:214–223, 2023. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.03.015.

Supriya Misra, Valerie W Jackson, Jeanette Chong, Karen Choe, Charisse Tay, Jazmine Wong, and Lawrence H Yang. Systematic review of cultural aspects of stigma and mental illness among racial and ethnic minority groups in the united states: implications for interventions. *American journal of community psychology*, 68(3-4):486–512, 2021.

Jihyung Moon, Hyunchang Cho, and Eunjeong L. Park. Revisiting round-trip translation for quality estimation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 91–104, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.11.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. Cross-lingual few-shot hate speech and offensive language detection using meta learning. *IEEE Access*, 10:14880–14896, 2022. doi: 10.1109/ACCESS.2022.3147588.

Thomas Mueller, Guillermo Pérez-Torró, and Marc Franco-Salvador. Few-shot learning with siamese networks and label tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8532–8545, 2022.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. L-HSAB: A Levantine Twitter dataset for hate speech and abusive language. In Sarah T. Roberts, Joel Tetreault, Vinodkumar Prabhakaran, and Zeerak Waseem, editors, *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3512. URL https://aclanthology.org/W19-3512.

Myriam Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on affective computing*, 5(2):101–111, 2014.

Simo Määttä. Linguistic and discursive properties of hate speech and speech facilitating the expression of hatred: Evidence from finnish and french online discussion boards. *Internet Pragmatics*, 6, 09 2023. doi: 10.1075/ip.00094.maa.

Pansy Nandwani and Rupali Verma . A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1):81, 2021. doi: 10.1007/s13278-021-00776-6.

Matthew L Newman, Carla J Groom, Lori D Handelman, and James W Pennebaker. Gender differences in language use: An analysis of 14,000 text samples. *Discourse processes*, 45(3):211–236, 2008.

Dong Nguyen, Dolf Trieschnigg, A. Seza Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska de Jong. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1950–1961, Dublin, Ireland, 2014. Dublin City University and Association for Computational Linguistics.

Eugène Albert Nida. *Componential Analysis of Meaning: An Introduction to Semantic Structures*. Mouton, The Hague, 1975.

John T. Nockleby. Hate speech. In Leonard W. Levy and Kenneth L. Karst et al., editors, *Encyclopedia of the American Constitution*, volume 1. Macmillan, New York, 2nd edition, 2000.

Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. An in-depth analysis of implicit and subtle hate speech messages. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th*

Conference of the European Chapter of the Association for Computational Linguistics, pages 1997–2013, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.147. URL https://aclanthology.org/2023.eacl-main.147.

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

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1017–1029, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.88. URL https://aclanthology.org/2023.acl-short.88.

Lidiia Ostyakova, Veronika Smilga, Kseniia Petukhova, Maria Molchanova, and Daniel Kornev. ChatGPT vs. crowdsourcing vs. experts: Annotating open-domain conversations with speech functions. In Svetlana Stoyanchev, Shafiq Joty, David Schlangen, Ondrej Dusek, Casey Kennington, and Malihe Alikhani, editors, *Proceedings of the 24th Annual Meeting of the Special Interest Group* *on Discourse and Dialogue*, pages 242–254, Prague, Czechia, September 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.sigdial-1.23. URL https://aclanthology.org/2023.sigdial-1.23.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-4009. URL https://aclanthology.org/N19-4009.

Pia Pachinger, Allan Hanbury, Julia Neidhardt, and Anna Planitzer. Toward disambiguating the definitions of abusive, offensive, toxic, and uncivil comments. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 107–113, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.c3nlp-1.11. URL https://aclanthology.org/2023.c3nlp-1.11.

Endang Wahyu Pamungkas and Viviana Patti. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-2051. URL https://aclanthology.org/P19-2051.

Jana Papcunová, Marcel Martončik, Daniela Fedáková, et al. Hate speech operationalization: a preliminary examination of hate speech indicators and their structure. *Complex Intelligent Systems*, 9:2827–2842, 2023. doi: 10.1007/s40747-021-00561-0. URL https://doi.org/10.1007/s40747-021-00561-0.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Bhikhu Parekh. Is there a case for banning hate speech? In Michael Herz and

Peter Molnar, editors, *The Content and Context of Hate Speech: Rethinking Regulation and Responses*, pages 37–56. Cambridge University Press, 2012.

Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1302. URL `https://aclanthology.org/D18-1302`.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. SemEval-2021 task 5: Toxic spans detection. In Alexis Palmer, Nathan Schneider, Natalie Schluter, Guy Emerson, Aurelie Herbelot, and Xiaodan Zhu, editors, *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.semeval-1.6. URL `https://aclanthology.org/2021.semeval-1.6`.

John Pavlopoulos, Leo Laugier, Alexandros Xenos, Jeffrey Sorensen, and Ion Androutsopoulos. From the detection of toxic spans in online discussions to the analysis of toxic-to-civil transfer. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3721–3734, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.259. URL `https://aclanthology.org/2022.acl-long.259`.

Jiaxin Pei and David Jurgens. When do annotator demographics matter? measuring the influence of annotator demographics with the POPQUORN dataset. In Jakob Prange and Annemarie Friedrich, editors, *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 252–265, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.law-1.25.

James W Pennebaker. The secret life of pronouns. *New Scientist*, 211(2828):42–45, 2011.

James W Pennebaker and Laura A King. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296, 1999.

James W Pennebaker and Lori D Stone. Words of wisdom: language use over the life span. *Journal of personality and social psychology*, 85(2):291, 2003.

James W Pennebaker, Tracy J Mayne, and Martha E Francis. Linguistic predictors of adaptive bereavement. *Journal of personality and social psychology*, 72(4): 863, 1997.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

Denis Peskov, Viktor Hangya, Jordan Boyd-Graber, and Alexander Fraser. Adapting entities across languages and cultures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3725–3750, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.315. URL https://aclanthology.org/2021.findings-emnlp.315.

Steven Pinker. *The Stuff of Thought: Language as a Window into Human Nature*. Viking, 2007.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. Offensive language identification in Greek. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.629.

Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In Yi-ling Chung, Paul R\"ottger, Debora Nozza, Zeerak Talat, and Aida Mostafazadeh Davani, editors, *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.woah-1.6. URL https://aclanthology.org/2023.woah-1.6.

Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350, 2001.

Fabio Poletto, Valerio Basile, Cristina Bosco, Viviana Patti, and Marco Antonio Stranisci. Annotating hate speech: Three schemes at comparison. In *Italian Conference on Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:204901524.

Francesca Poletto, Valerio Basile, Manuela Sanguinetti, et al. Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*, 55(2):477–523, 2021. doi: 10.1007/s10579-020-09502-8. URL https://doi.org/10.1007/s10579-020-09502-8.

Marco Polignano, Pierpaolo Basile, Marco Degemmis, Giovanni Semeraro, and Valerio Basile. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *Italian Conference on Computational Linguistics*, 2019a. URL https://api.semanticscholar.org/CorpusID:204914950.

Marco Polignano, Pierpaolo Basile, Marco De Gemmis, and Giovanni Semeraro. Hate speech detection through alberto italian language understanding model. In *NL4AI@AI*IA*, 2019b. URL https://api.semanticscholar.org/CorpusID:209443031.

Kimberly A Quinn and C Neil Macrae. Categorizing others: the dynamics of person construal. *Journal of personality and social psychology*, 88(3):467, 2005.

Gil Ramos, Fernando Batista, Ricardo Ribeiro, Pedro Fialho, Sérgio Moro, António Fonseca, Rita Guerra, Paula Carvalho, Catarina Marques, and Cláudia Silva. A comprehensive review on automatic hate speech detection in the age of the transformer. *Social Network Analysis and Mining*, 14(204), 2024. doi: 10.1007/s13278-024-01361-3. URL https://doi.org/10.1007/s13278-024-01361-3.

Mattia Retta. A pragmatic and discourse analysis of hate words on social media. *Internet Pragmatics*, 6(2):197–218, 2023. ISSN 2542-3851. doi: https://doi.org/10.1075/ip.00096.ret. URL https://www.jbe-platform.com/content/journals/10.1075/ip.00096.ret.

Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA ’21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380959.

doi: 10.1145/3411763.3451760. URL https://doi.org/10.1145/3411763.3451760.

Howard Rheingold. *The Virtual Community: Homesteading on the Electronic Frontier*. The MIT Press, 10 2000.

Giulia Rizzi, Michele Fontana, and Elisabetta Fersini. Perspectives on hate: General vs. domain-specific models. In Gavin Abercrombie, Valerio Basile, Davide Bernadi, Shiran Dudy, Simona Frenda, Lucy Havens, and Sara Tonelli, editors, *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 78–83, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.nlperspectives-1.8.

Gal Ron, Effi Levi, Odelia Oshri, and Shaul Shenhav. Factoring hate speech: A new annotation framework to study hate speech in social media. In Yi-ling Chung, Paul R\"ottger, Debora Nozza, Zeerak Talat, and Aida Mostafazadeh Davani, editors, *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 215–220, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.woah-1.21. URL https://aclanthology.org/2023.woah-1.21.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. Two contrasting data annotation paradigms for subjective NLP tasks. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.13. URL https://aclanthology.org/2022.naacl-main.13.

Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. Probing LLMs for hate speech detection: strengths and vulnerabilities. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6116–6128, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.407. URL https://aclanthology.org/2023.findings-emnlp.407.

Federico Ruggeri, Francesco Antici, Andrea Galassi, Katerina Korre, Arianna Muti, and Alberto Barrón-Cedeño. On the definition of prescriptive annotation guidelines for language-agnostic subjectivity detection. In *Proceedings of the Text2Story'23 Workshop*, 2023.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France, June 2022a. European Language Resources Association. URL https://aclanthology.org/2022.nlperspectives-1.11.

Pratik S. Sachdeva, Renata Barreto, Claudia von Vacano, and Chris J. Kennedy. Assessing annotator identity sensitivity via item response theory: A case study in a hate speech corpus. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1585–1603, New York, NY, USA, 2022b. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533216.

Hind Saleh, Areej Alhothali, and Kawthar Moria. Detection of hate speech using bert and hate speech word embedding with deep model. *Applied Artificial Intelligence*, 37(1), 2023. doi: 10.1080/08839514.2023.2166719. URL https://doi.org/10.1080/08839514.2023.2166719.

Maurizio Sanguinetti, Giacomo Comandini, Elia di Nuovo, Simone Frenda, Maria Stranisci, Cristina Bosco, and Irene Russo. Haspeede 2 @ evalita2020: Overview of the evalita 2020 hate speech detection task. In Valerio Basile, Danilo Croce, Maria Maro, and Lucia C. Passaro, editors, *EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020: Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian Final Workshop*, Torino, 2020. Accademia University Press. doi: 10.4000/books.aaccademia.6897.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1163. URL https://aclanthology.org/P19-1163.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.486. URL https://aclanthology.org/2020.acl-main.486.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.431.

Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1101. URL https://aclanthology.org/W17-1101/.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.117. URL https://aclanthology.org/2020.emnlp-main.117.

Anders Søgaard. Should we ban English NLP for a year? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5260, Abu Dhabi, United Arab Emirates, December 2022. Association for

Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.351. URL https://aclanthology.org/2022.emnlp-main.351.

Sanja Štajner. Exploring reliability of gold labels for emotion detection in twitter. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1350–1359, online, 2021.

Sanja Štajner, Seren Yenikent, and Marc Franco-Salvador. Five psycholinguistic characteristics for better interaction with users. In *2021 8th International Conference on Behavioral and Social Computing (BESC)*, pages 1–7. IEEE, 2021.

Lukas Stappen, Fabian Brunn, and Björn Schuller. Cross-lingual Zero- and Few-shot Hate Speech Detection Utilising Frozen Transformer Language Models and AXEL. *arXiv e-prints*, art. arXiv:2004.13850, April 2020. doi: 10.48550/arXiv.2004.13850.

Carlo Strapparava and Rada Mihalcea. *Annotating and Identifying Emotions in Text*, pages 21–38. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. ISBN 978-3-642-14000-6. doi: 10.1007/978-3-642-14000-6_2.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, Manfred Klenner, and et al. Overview of germeval task 2, 2019 shared task on the identification of offensive language. In *Proceedings of the GermEval 2019*, 2019.

Kaili Sun, Xudong Luo, and Michael Y. Luo. A survey of pretrained language models. In Gerard Memmi, Baijian Yang, Linghe Kong, Tianwei Zhang, and Meikang Qiu, editors, *Knowledge Science, Engineering and Management*, pages 442–456, Cham, 2022. Springer International Publishing. ISBN 978-3-031-10986-7.

Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/K19-1088. URL https://aclanthology.org/K19-1088.

Zeerak Talat, James Thorne, and Joachim Bingel. *Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection*, pages 29–55.

2018. doi: 10.1007/978-3-319-78583-7_3. URL https://app.dimensions.ai/details/publication/pub.1105734529.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation and synthesis: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.54. URL https://aclanthology.org/2024.emnlp-main.54.

Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

Teodor Tita and Arkaitz Zubiaga. Cross-lingual hate speech detection using transformer models. *CoRR*, abs/2111.00981, 2021. URL https://arxiv.org/abs/2111.00981.

Ashish Tiwari. Chapter 2 - supervised learning: From theory to applications. In Rajiv Pandey, Sunil Kumar Khatri, Neeraj kumar Singh, and Parul Verma, editors, *Artificial Intelligence and Machine Learning for EDGE Computing*, pages 23–32. Academic Press, 2022. ISBN 978-0-12-824054-0. doi: https://doi.org/10.1016/B978-0-12-824054-0.00026-5. URL https://www.sciencedirect.com/science/article/pii/B9780128240540000265.

Joshua Aaron Tucker, Andrew Guess, Pablo Barbera, Cristian Vaccari, Alexandra Siegel, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. Social media, political polarization, and political disinformation: A review of the scientific literature. *SSRN Electronic Journal*, 2018.

Joshua M Tybur, Debra Lieberman, and Vladas Griskevicius. Microbes, mating, and morality: individual differences in three functional domains of disgust. *Journal of personality and social psychology*, 97(1):103, 2009.

Nicola Ueffing, Gholamreza Haffari, and Anoop Sarkar. Transductive learning for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 25–32, Prague, Czech

Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/P07-1004.

UN General Assembly. International covenant on civil and political rights. Available from the United Nations. Adopted by the General Assembly.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 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, 2017. Curran Associates Inc. ISBN 9781510860964.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. Learning from the worst: Dynamically generated datasets to improve online hate detection. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.132. URL https://aclanthology.org/2021.acl-long.132.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *ACL*, 2021b.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. Multilingual is not enough: Bert for finnish, 2019. URL https://arxiv.org/abs/1912.07076.

F Schulz Von Thun. Miteinander reden 1-allgemeine psychologie der kommunikation. *Frank Naumann: Die Kunst des Small Talk "Leicht ins Gespräch kommen. Helga Schuler:„Erfolgreiches Telefonieren "Harry Holzheu:„Natürliche Rhetorik*, 1981.

Nina Wacholder, Smaranda Muresan, Debanjan Ghosh, and Mark Aakhus. Annotating multiparty discourse: Challenges for agreement metrics. In Lori Levin and Manfred Stede, editors, *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 120–128, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University. doi: 10.3115/v1/W14-4918. URL https://aclanthology.org/W14-4918.

Mark Walters, Rupert Brown, and Susann Wiedlitzka. Causes and motivations of hate crime. *Equality and Human Rights Commission research report*, 102, 2016.

Zeerak Waseem. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In David Bamman, A. Seza Doğruöz, Jacob Eisenstein, Dirk Hovy, David Jurgens, Brendan O'Connor, Alice Oh, Oren Tsur, and Svitlana Volkova, editors, *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-5618. URL https://aclanthology.org/W16-5618.

Gregor Wiedemann, Seid Muhie Yimam, and Chris Biemann. UHH-LT at SemEval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1638–1644, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.213. URL https://aclanthology.org/2020.semeval-1.213.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. Overview of the germeval 2018 shared task on the identification of offensive language. Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), Vienna, Austria – September 21, 2018, pages 1 – 10. Austrian Academy of Sciences, Vienna, Austria, 2019. ISBN 978-3-7001-8435-5. URL https://nbn-resolving.org/urn:nbn:de:bsz:mh39-84935.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. Implicitly abusive language – what does it actually look like and why are we not getting there? In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.48. URL https://aclanthology.org/2021.naacl-main.48/.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale, 2017.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. HARE: Explainable hate speech detection with step-by-step reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5490–5505, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.365. URL https://aclanthology.org/2023.findings-emnlp.365.

Wenjie Yin and Arkaitz Zubiaga. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7, 2021.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2010. URL https://aclanthology.org/S19-2010.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.188. URL https://aclanthology.org/2020.semeval-1.188.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15(2), February 2024. ISSN 2157-6904. doi: 10.1145/3639372. URL https://doi.org/10.1145/3639372.

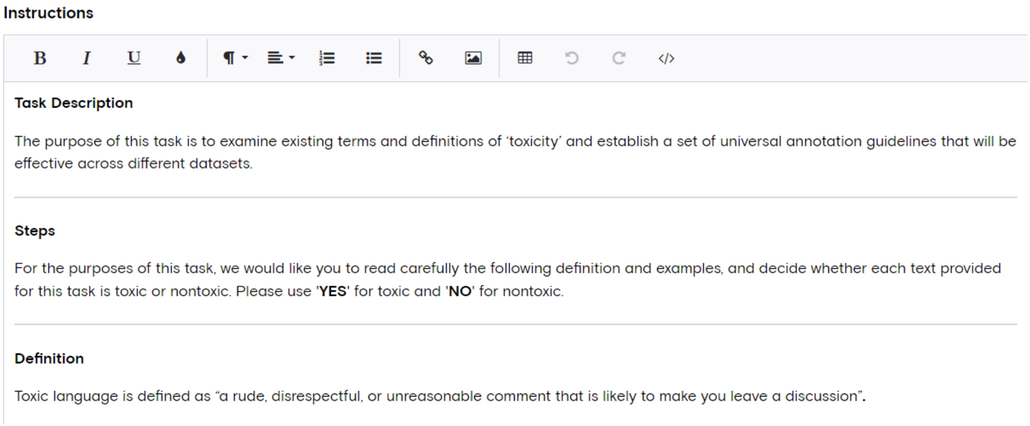
Haris Bin Zia, Ignacio Castro, Arkaitz Zubiaga, and Gareth Tyson. Improving zero-shot cross-lingual hate speech detection with pseudo-label finetuning of transformer language models. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):1435–1439, May 2022. doi: 10.1609/icwsm.v16i1.19402. URL https://ojs.aaai.org/index.php/ICWSM/article/view/19402.

Frederike Zufall, Marius Hamacher, Katharina Kloppenborg, and Torsten Zesch. A legal approach to hate speech – operationalizing the EU's legal framework against the expression of hatred as an NLP task. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 53–64, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.nllp-1.5. URL https://aclanthology.org/2022.nllp-1.5.

Appendix A

Annotation Instructions and Interface

Figures A.1 and A.2 of this Appendix present the instructions (only for toxicity shown) and the interface the annotators were provided with during their re-annotation tasks.



Instructions

Task Description

The purpose of this task is to examine existing terms and definitions of 'toxicity' and establish a set of universal annotation guidelines that will be effective across different datasets.

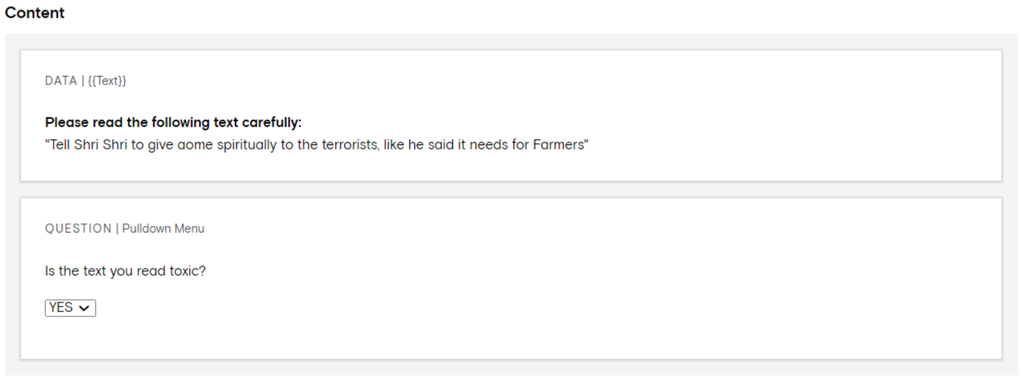
Steps

For the purposes of this task, we would like you to read carefully the following definition and examples, and decide whether each text provided for this task is toxic or nontoxic. Please use '**YES**' for toxic and '**NO**' for nontoxic.

Definition

Toxic language is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion".

Figure A.1: Re-annotation instructions, using the term and definition of Toxicity.



Content

DATA | {{Text}}

Please read the following text carefully:
"Tell Shri Shri to give aome spiritually to the terrorists, like he said it needs for Farmers"

QUESTION | Pulldown Menu

Is the text you read toxic?

YES

Figure A.2: Interface for re-annotation.

Appendix B

Laws, Prompts, and Further Scores

B.1 Laws

In this appendix we reproduce the articles that describe prosecutable hate speech in the three countries .

B.1.1 Greece

Law No 927/1797 on Punishing Acts or Activities aimed at Racial Discrimination.

Article 1: Incitement to violence or hatred.
(1) “Anyone, who publicly incites, provokes, or stirs up, either orally or through the press, the Internet, or any other means, acts of violence or hatred against a person or group of persons or a member of such a group defined by reference to race, color, religion, descent or national or ethnic origin, sexual orientation, gender identity, or disability, in a manner that endangers the public order and exposes the life, physical integrity, and freedom of persons defined above to danger, will be punished by imprisonment of from three months to three years and a fine of Euros 5,000 to 20,000.
(2) Anyone, who publicly incites, provokes, or stirs, either orally or through the press, the Internet, or any other means, acts of destruction against the assets of a person or group of persons defined by reference to race, color, religion, descent or national or ethnic origin, sexual orientation, gender identity, or disability, in a manner that endangers the public order and exposes the life, physical integrity, and

freedom of persons defined above to danger, will be punished by imprisonment of from three months to three years and a fine of Euros 5,000 to 20,000.

B.1.2 Italy

Criminal code. Article 604 bis: Propaganda and incitement to commit crimes for reasons of racial, ethnic and religious discrimination.

I. Unless the fact constitutes a more serious offence, the following are punished: a) with imprisonment of up to one year and six months or with a fine of up to 6,000 euros for propaganda based on ideas of superiority or ideas of racial or ethnic hatred, or propaganda which instigates to commit or commits acts of discrimination on the grounds of racial, ethnic, national or religious grounds; b) Imprisonment of six months to four years for a person who in any way, instigates to commit or commits violence or acts of provocation to violence for racial, ethnic, national or religious reasons.

Law 654/1975 on the Ratification and Execution of the International Convention on the Elimination of All Forms of Racial Discrimination:

Article 3: I. Unless the fact constitutes a more serious offence, the following are punished: a) with imprisonment of up to one year and six months or with a fine of up to 6,000 euros for propaganda based on ideas of superiority or ideas of racial or ethnic hatred, or propaganda which instigates to commit or commits acts of discrimination on the grounds of racial, ethnic, national or religious grounds; b) Imprisonment of six months to four years for a person who in any way, instigates to commit or commits violence or acts of provocation to violence for racial, ethnic, national or religious. Law 167/2017: Full Implementation of EU Framework Decision 2008/913/HAD This law, amongst others modifies Law 654/1975: The main relevant Italian Law is Law 205/1993 which makes it a crime to “propagate ideas based on racial superiority or racial or ethnic hatred, or to instigate to commit or commit acts of discrimination for racial, ethnic, national or religious motives.” The law also punishes those who “instigate in any way or commit violence or acts of provocation to violence for racist, ethnic, national or religious motives.”

B.1.3 UK

Public Order Act 1986 Part III Racial Hatred

Acts intended or likely to stir up racial hatred Section 18 Use of words or behaviour or display of written material. (1) A person who uses threatening, abusive or insulting words or behaviour, or displays any written material which is threatening, abusive or insulting, is guilty of an offence if— (a)he intends thereby to stir up racial hatred, or (b)having regard to all the circumstances racial hatred is likely to be stirred up thereby. (2) An offence under this section may be committed in a public or a private place, except that no offence is committed where the words or behaviour are used, or the written material is displayed, by a person inside a dwelling and are not heard or seen except by other persons in that or another dwelling. (4) In proceedings for an offence under this section it is a defence for the accused to prove that he was inside a dwelling and had no reason to believe that the words or behaviour used, or the written material displayed, would be heard or seen by a person outside that or any other dwelling. (5) A person who is not shown to have intended to stir up racial hatred is not guilty of an offence under this section if he did not intend his words or behaviour, or the written material, to be, and was not aware that it might be, threatening, abusive or insulting. (6) This section does not apply to words or behaviour used, or written material displayed, solely for the purpose of being included in a programme [included in a programme service]. Section 19 Publishing or distributing written material. (1)A person who publishes or distributes written material which is threatening, abusive or insulting is guilty of an offence if— (a)he intends thereby to stir up racial hatred, or (b)having regard to all the circumstances racial hatred is likely to be stirred up thereby. (2) In proceedings for an offence under this section it is a defence for an accused who is not shown to have intended to stir up racial hatred to prove that he was not aware of the content of the material and did not suspect, and had no reason to suspect, that it was threatening, abusive or insulting. (3) References in this Part to the publication or distribution of written material are to its publication or distribution to the public or a section of the public. Section 20 Public performance of play. (1) If a public performance of a play is given which involves the use of threatening, abusive or insulting words or behaviour, any person who presents or directs the performance is guilty of an offence if— (a)he intends thereby to stir up racial hatred, or (b)having regard to all the circumstances (and, in particular, taking the performance as a whole) racial hatred is likely to be stirred up thereby. (2) If a person presenting or directing the performance is not shown to have intended to stir up racial hatred, it is a defence for him to prove— (a) that he did not know and had no reason to suspect that the performance would involve the use of the offending words or behaviour, or (b) that he did not know and had no reason to suspect that the offending words or behaviour were threatening, abusive or insulting, or (c) that he did not know and had no reason to suspect that the circumstances in which the performance would be given would be such that racial hatred would be likely to be stirred up. (3) This section does not apply to a performance given solely or primarily for one or more of the following purposes— (a) rehearsal, (b) making a recording of the performance, or (c) enabling the performance to be [included in a programme service]; but if it is proved that the performance was attended by persons other than those directly connected with the giving of the performance or the doing in relation to it of the things mentioned in paragraph (b) or (c), the performance shall, unless the contrary is shown, be taken not to have been given solely or primarily for the purposes mentioned above. (4) For the purposes of this section— (a)a person shall not be treated as presenting a performance of a play by reason only of his taking part in it as a performer, (b) a person taking part as a performer in a performance directed by another shall be treated as a person who directed the performance if without reasonable excuse he performs otherwise than in accordance with that person's direction, and (c) a person shall be taken to have directed a performance of a play given under his direction notwithstanding that he was not present during the performance; and a person shall not be treated as aiding or abetting the commission of an offence under this section by reason only of his taking part in a performance as a performer. (5) In this section "play" and "public performance" have the same meaning as in the Theatres Act 1968. (6) The following provisions of the Theatres Act 1968 apply in relation to an offence under this section as they apply to an offence under section 2 of that Act— section 9 (script as evidence of what was performed), section 10 (power to make copies of script), section 15 (powers of entry and inspection). Section 21 Distributing, showing or playing a recording. (1) A person who distributes, or shows or plays, a recording of visual images or sounds which are threatening, abusive or insulting is guilty of an offence if— (a) he intends thereby to stir up racial hatred, or (b) having regard to all the circumstances racial hatred is likely to be stirred up thereby. (2) In this Part "recording" means any record from which visual images or sounds may, by any means, be reproduced; and references to the distribution, showing or playing of a recording are to its distribution, showing or playing of a recording are to its distribution, showing or playing to the public or a section of the public. (3) In proceedings for an offence under this section it is a defence for an accused who is not shown to have intended to stir up racial hatred to prove that he was not aware of the content of the recording and did not suspect, and had no reason to suspect, that it was threatening, abusive or insulting. (4) This section does not apply to the showing or playing of a recording solely for the purpose of enabling the recording to be included in a programme service. Section 22 Broadcasting or including programme in cable programme service. (1) If a programme involving threatening, abusive or insulting visual images or sounds is included in a programme service], each of the persons mentioned in subsection (2) is guilty of an offence if— (a)he intends thereby to stir up racial hatred, or (b)having regard to all the circumstances racial hatred is likely to be stirred up thereby. (2) The persons are— (a)the person providing the programme service, (b)any person by whom the programme is produced or directed, and (c)any person by whom offending words or behaviour are used. (3) If the person providing the service, or a person by whom the programme was produced or directed, is not shown to have intended to stir up racial hatred, it is a defence for him to prove that— (a)he did not know and had no reason to suspect that the programme would involve the offending material, and (b)having regard to the circumstances in which the programme was [included in a programme service], it was not reasonably practicable for him to secure the removal of the material. (4) It is a defence for a person by whom the programme was produced or directed who is not shown to have intended to stir up racial hatred to prove that he did not know and had no reason to suspect— (a)that the programme would be [included in a programme service], or (b)that the circumstances in which the programme would be so included would be such that racial hatred would be likely to be stirred up. (5) It is a defence for a person by whom offending words or behaviour were used and who is not shown to have intended to stir up racial hatred to prove that he did not know and had no reason to suspect— (a)that a programme involving the use of the offending material would be [included in a programme service], or (b)that the circumstances in which a programme involving the use of the offending material would be so included, or in which a programme . . . so included would involve the use of the offending material, would be such that racial hatred would be likely to be stirred up. (6) A person who is not shown to have intended to stir up racial hatred is not guilty of an offence under this section if he did not know, and had no reason to suspect, that the offending material was threatening, abusive or insulting. Racially inflammatory material Section 23 Possession of racially inflammatory material. (1) A person who has in his possession written material which is threatening, abusive or insulting, or a recording of visual images or sounds which are threatening, abusive or insulting, with a view to— (a)in the case of written material, its being displayed, published, distributed, [or included in a cable programme service], whether by himself or another, or (b)in the case of a recording, its being distributed, shown, played, [or included in a cable programme service], whether by himself or another, is guilty of an offence if he intends racial hatred to be stirred up thereby or, having regard to all the circumstances, racial hatred is likely to be stirred up thereby. (2) For this purpose regard shall be had to such display, publication, distribution, showing, playing, [or inclusion in a programme service] as he has, or it may reasonably be inferred that he has, in view. (3) In proceedings for an offence under this section it is a defence for an accused who is not shown to have intended to stir up racial hatred to prove that he was not aware of the content of the written material or recording and did not suspect, and had no reason to suspect, that it was threatening, abusive or insulting. Part 3A Hatred against persons on religious grounds Acts intended to stir up religious hatred Section 29B: Use of words or behaviour or display of written material (1) A person who uses threatening words or behaviour, or displays any written material which is threatening, is guilty of an offence if he intends thereby to stir up religious hatred. (2) An offence under this section may be committed in a public or a private place, except that no offence is committed where the words or behaviour are used, or the written material is displayed, by a person inside a dwelling and are not heard or seen except by other persons in that or another dwelling. (3) A constable may arrest without warrant anyone he reasonably suspects is committing an offence under this section. (4) In proceedings for an offence under this section it is a defence for the accused to prove that he was inside a dwelling and had no reason to believe that the words or behaviour used, or the written material displayed, would be heard or seen by a person outside that or any other dwelling. (5) This section does not apply to words or behaviour used, or written material displayed, solely for the purpose of being included in a programme service. 29C Publishing or distributing written material (1) A person who publishes or distributes written material which is threatening is guilty of an offence if he intends thereby to stir up religious hatred. (2) References in this Part to the publication or distribution of written material are to its publication or distribution to the public or a section of the public. Section 29D: Public performance of play (1) If a public performance of a play is given which involves the use of threatening words or behaviour, any person who presents or directs the performance is guilty of an offence if he intends thereby to stir up religious hatred. (2) This section does not apply to a performance given solely or primarily for one or more of the following purposes— (a)rehearsal, (b)making a recording of the performance, or (c)enabling the performance to be included in a programme service; but if it is proved that the performance was attended by persons other than those directly connected with the giving of the performance or the doing in relation to it of the things mentioned in paragraph (b) or (c), the performance shall, unless the contrary is shown, be taken not to have been given solely or primarily for the purpose mentioned above. (3) For the purposes of this section— (a)a person shall not be treated as presenting a performance of a play by reason only of his taking part in it as a performer, (b)a person taking part as a performer in a performance directed by another shall be treated as a person who directed the performance if without reasonable excuse he performs otherwise than in accordance with that person's direction, and (c)a person shall be taken to have directed a performance of a play given under his direction notwithstanding that he was not present during the performance; and a person shall not be treated as aiding or abetting the commission of an offence under this section by reason only of his taking part in a performance as a performer. (4) In this section "play" and "public performance" have the same meaning as in the Theatres Act 1968. (5) The following provisions of the Theatres Act 1968 apply in relation to an offence under this section as they apply to an offence under section 2 of that Act— section 9 (script as evidence of what was performed), section 10 (power to make copies of script), section 15 (powers of entry and inspection). Section 29E Distributing, showing or playing a recording (1) A person who distributes, or shows or plays, a recording of visual images or sounds which are threatening is guilty of an offence if he intends thereby to stir up religious hatred. (2) In this Part "recording" means any record from which visual images or sounds may, by any means, be reproduced; and references to the distribution, showing or playing of a recording are to its distribution, showing or playing to the public or a section of the public. (3) This section does not apply to the showing or playing of a recording solely for the purpose of enabling the recording to be included in a programme service. Section 29F Broadcasting or including programme in programme service (1) If a programme involving threatening visual images or sounds is included in a programme service, each of the persons mentioned in subsection (2) is guilty of an offence if he intends thereby to stir up religious hatred. (2) The persons are— (a)the person providing the programme service, (b)any person by whom the programme is produced or directed, and (c)any person by whom offending words or behaviour are used. Inflammatory material Section 29G Possession of inflammatory material (1) A person who has in his possession written material which is threatening, or a recording of visual images or sounds which are threatening, with a view to— (a)in the case of written material, its being displayed, published, distributed, or included in a programme service whether by himself or another, or (b)in the case of a recording, its being distributed, shown, played, or included in a programme service, whether by himself or another, is guilty of an offence if he intends religious hatred to be stirred up thereby. (2) For this purpose regard shall be had to such display, publication, distribution, showing, playing, or inclusion in a programme service as he has, or it may reasonably be inferred that he has, in view. Section 29H: Powers of entry and search (1) If in England and Wales a justice of the peace is satisfied by information on oath laid by a constable that there are reasonable grounds for suspecting that a person has possession of written material or a recording in contravention of section 29G, the justice may issue a warrant under his hand authorising any constable to enter and search the premises where it is suspected the material or recording is situated. (2) If in Scotland a sheriff or justice of the peace is satisfied by evidence on oath that there are reasonable grounds for suspecting that a person has possession of written material or a recording in contravention of section 29G, the sheriff or justice may issue a warrant authorising any constable to enter and search the premises where it is suspected the material or recording is situated. (3) A constable entering or searching premises in pursuance of a warrant issued under this section may use reasonable force if necessary. (4) In this section "premises" means any place and, in particular, includes— (a)any vehicle, vessel, aircraft or hovercraft, (b)any offshore installation as defined in section 12 of the Mineral Workings (Offshore Installations) Act 1971, and (c)any tent or movable structure. Section 29I: Power to order forfeiture (1) A court by or before which a person is convicted of— (a)an offence under section 29B relating to the display of written material, or (b)an offence under section 29C, 29E or 29G, shall order to be forfeited any written material or recording produced to the court and shown to its satisfaction to be written material or a recording to which the offence relates. (2) An order made under this section shall not take effect— (a)in the case of an order made in proceedings in England and Wales, until the expiry of the ordinary time within which an appeal may be instituted or, where an appeal is duly instituted, until it is finally decided or abandoned; (b)in the case of an order made in proceedings in Scotland, until the expiration of the time within which, by virtue of any statute, an appeal may be instituted or, where such an appeal is duly instituted, until the appeal is finally decided or abandoned. (3) For the purposes of subsection (2)(a)— (a)an application for a case stated or for leave to appeal shall be treated as the institution of an appeal, and (b)where a decision on appeal is subject to a further appeal, the appeal is not finally determined until the expiry of the ordinary time within which a further appeal may be instituted or, where a further appeal is duly instituted, until the further appeal is finally decided or abandoned. (4) For the purposes of subsection (2)(b) the lodging of an application for a stated case or note of appeal against sentence shall be treated as the institution of an appeal. Section 29J Protection of freedom of expression Nothing in this Part shall be read or given effect in a way which prohibits or restricts discussion, criticism or expressions of antipathy, dislike, ridicule, insult or abuse of particular religions or the beliefs or practices of their adherents, or of any other belief system or the beliefs or practices of its adherents, or proselytising or urging adherents of a different religion or belief system to cease practising their religion or belief system.

B.2 Annotation Guidelines

The guidelines provided for the annotation are reproduced next.

Hate speech laws annotation

You are asked to annotate 100 instances according to 3 different hate speech laws. Please find the sentences in the separate excel file. All of these sentences are originally annotated as hate speech; however, we want to know how many of these can be prosecutable hate speech under three different laws: Greek, Italian, UK. The laws are taken from this website. You can find the laws down below. Your job is to annotate in the designated column whether an instance is **prosecutable** hate speech by assigning the value 1 or not by assigning the value 0. Feel free to use the comments column for any thoughts. If you have any questions contact. . .

			Legal context		
id	**Text**	**Greek**	**Italian**	**UK**	**comments**
Id1	Instance1	0	1	1	
Id2	Instance2	1	0	1	
. . .					

B.3 Prompts

0-shot_g

Is the following sentence prosecutable hate speech according? Reply with JUST a number:(0) not prosecutable (1) unlikely prosecutable (2) likely prosecutable (3) surely prosecutable. [sentence]

0-shot w/Law

Is the following sentence prosecutable hate speech according to the [country name] law? Reply with JUST a number:(0) not prosecutable (1) unlikely prosecutable (2) likely prosecutable (3) surely prosecutable. [country law] [sentence]

few-shot & LOOCV

Is the following sentence prosecutable hate speech according to the examples? Reply with JUST a number:(0) not prosecutable (1) unlikely prosecutable (2) likely prosecutable (3) surely prosecutable.[examples][sentence]

few-shot & LOOCV w/Law

Is the following sentence prosecutable hate speech according to the examples and the law? Reply with JUST a number:(0) not prosecutable (1) unlikely prosecutable (2) likely prosecutable (3) surely prosecutable.[law] [examples][sentence]

B.4 Multilabel Error Scores and Multiclass F1 Scores

	HateBERT			**DehateBERT**			**HateRoBERTa**			**LegalBERT**		
Country	Mic	Mac	Weight	Mic	Mac	Weight	Mic	Mac	Weight	Mic	Mac	Weight
Greece	0.69	0.47	0.66	0.55	0.38	0.47	0.58	0.37	0.54	0.50	0.29	0.46
Italy	0.75	0.47	0.70	0.64	0.52	0.34	0.69	0.53	0.66	0.64	0.37	0.60
UK	0.72	0.53	0.67	0.67	0.37	0.60	0.69	0.45	0.66	0.67	0.37	0.63

Table B.1: F_1 scores of PLMs.

	HateBERT			**DehateBERT**			**HateRoBERTa**			**LegalBERT**		
Country	Mic	Mac	Weight	Mic	Mac	Weight	Mic	Mac	Weight	Mic	Mac	Weight
Greece	0.32	0.32	0.19	0.14	0.11	0.12	0.42	0.24	0.36	0.12	0.08	0.12
Italy	0.20	0.12	0.20	0.13	0.10	0.11	0.44	0.21	0.40	0.13	0.07	0.08
UK	0.24	0.15	0.27	0.17	0.11	0.18	0.59	0.19	0.44	0.11	0.06	0.08

Table B.2: F_1 scores of PLMs per class when trained on the silver labels.

	0-shot			**0-shot w/Law**			**LOOCV few-shot**			**LOOCV few-shot w/Law**		
Country	Mic	Mac	Weight	Mic	Mac	Weight	Mic	Mac	Weight	Mic	Mac	Weight
Greece	0.26	0.26	0.24	0.37	0.28	0.34	0.52	0.31	0.58	0.53	0.35	0.57
Italy	0.25	0.24	0.24	0.37	0.29	0.31	0.59	0.28	0.70	0.56	0.27	0.64
UK	0.23	0.20	0.21	0.21	0.18	0.19	0.61	0.30	0.69	0.55	0.37	0.56

Table B.3: F_1 scores per class for Qwen2 for 0-shot and LOOCV settings.

	0-shot			**0-shot w/Law**			**LOOCV few-shot**			**LOOCV few-shot w/Law**		
Country	Mic	Mac	Weight	Mic	Mac	Weight	Mic	Mac	Weight	Mic	Mac	Weight
Greece	0.24	0.22	0.24	0.16	0.15	0.13	0.31	0.24	0.31	0.24	0.16	0.26
Italy	0.13	0.13	0.13	0.16	0.13	0.17	0.37	0.32	0.32	0.33	0.29	0.30
UK	0.20	0.21	0.18	0.28	0.22	0.26	0.35	0.33	0.32	0.32	0.30	0.30

Table B.4: F_1 scores per class for Llama3 for 0-shot and LOOCV settings.

	4-shot			**8-shot**			**12-shot**		
Country	Mic	Mac	Weight	Mic	Mac	Weight	Mic	Mac	Weight
Greece	0.37	0.29	0.36	0.41	0.25	0.45	0.39	0.32	0.43
Italy	0.41	0.30	0.39	0.46	0.33	0.45	0.59	0.40	0.62
UK	0.47	0.36	0.43	0.48	0.22	0.51	0.52	0.34	0.54

Table B.5: F_1 scores of Qwen2 in the few-shot setting without law.

	4-shot			**8-shot**			**12-shot**		
Country	Mic	Mac	Weight	Mic	Mac	Weight	Mic	Mac	Weight
Greece	0.35	0.27	0.31	0.30	0.26	0.26	0.34	0.29	0.32
Italy	0.33	0.27	0.30	0.28	0.22	0.24	0.28	0.21	0.26
UK	0.41	0.35	0.39	0.32	0.31	0.34	0.34	0.30	0.29

Table B.6: F_1 scores of Llama3 in the few-shot setting without law.

	4-shot			**8-shot**			**12-shot**		
Country	Mic	Mac	Weight	Mic	Mac	Weight	Mic	Mac	Weight
Greece	0.42	0.33	0.41	0.41	0.34	0.44	0.38	0.29	0.39
Italy	0.59	0.39	0.56	0.53	0.37	0.53	0.58	0.45	0.58
UK	0.45	0.35	0.40	0.42	0.28	0.37	0.51	0.37	0.48

Table B.7: F_1 scores of Qwen2 in the few-shot setting with law.

	4-shot			**8-shot**			**12-shot**		
Country	Mic	Mac	Weight	Mic	Mac	Weight	Mic	Mac	Weight
Greece	0.29	0.26	0.26	0.34	0.31	0.31	0.26	0.25	0.26
Italy	0.33	0.25	0.30	0.30	0.28	0.28	0.34	0.30	0.30
UK	0.37	0.32	0.36	0.32	0.27	0.29	0.29	0.25	0.30

Table B.8: F_1 scores of Llama3 in the few-shot setting with law.

B.5 Confusion Matrices

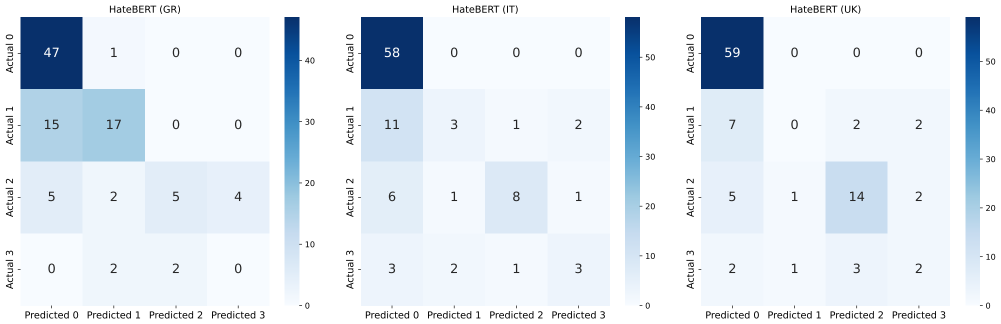


Figure B.1: HateBERT confusion matrix.

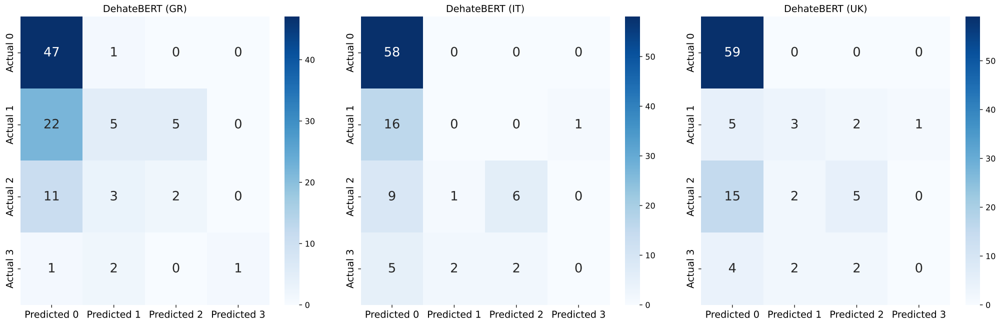


Figure B.2: DehateBERT confusion matrix.

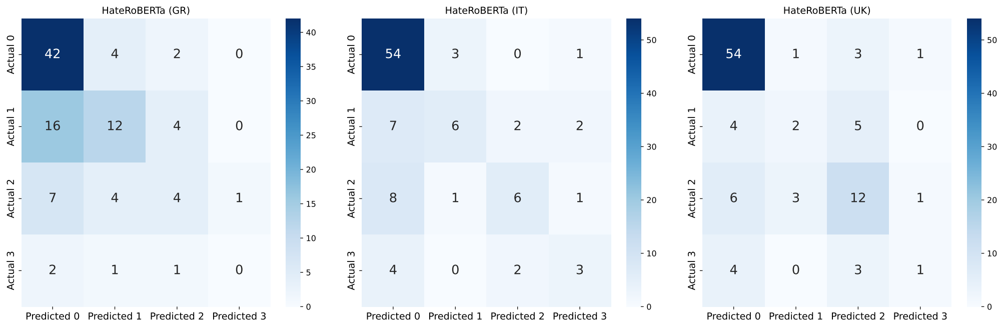


Figure B.3: HateRoBERTa confusion matrix.

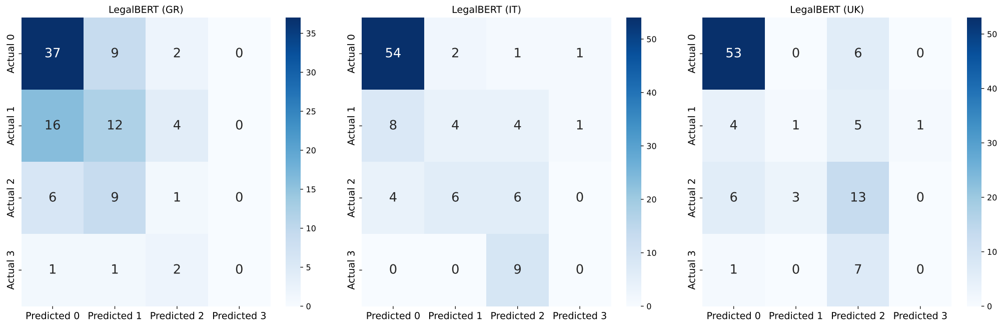


Figure B.4: LegalBERT confusion matrix.

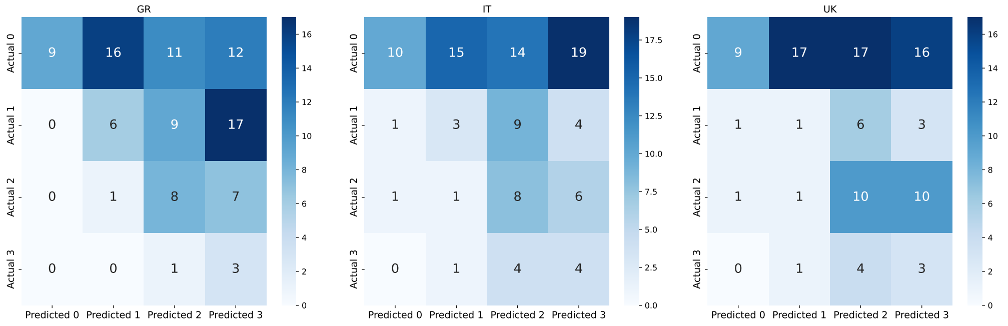


Figure B.5: Qwen2 0-shot w/o Law confusion matrix.

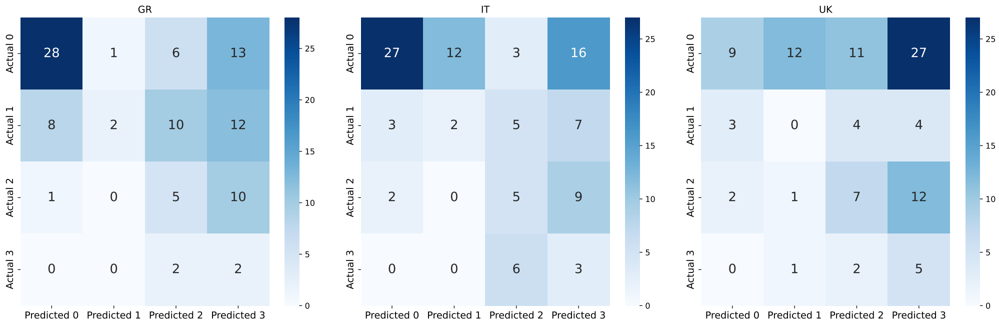


Figure B.6: Qwen2 0-shot w/ Law confusion matrix.

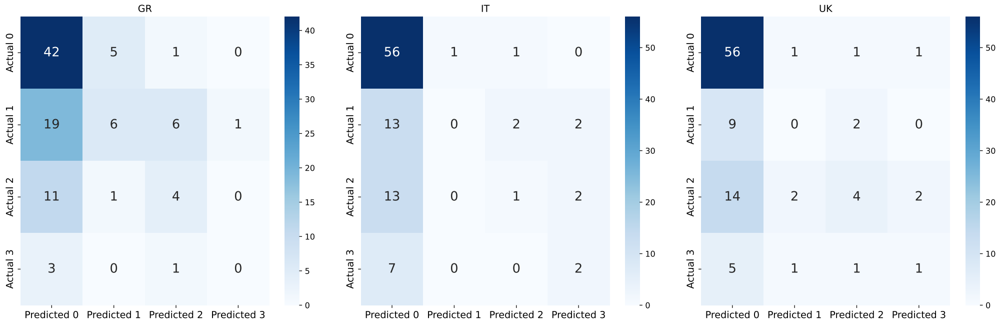


Figure B.7: Qwen2 LOOCV w/o Law confusion matrix.

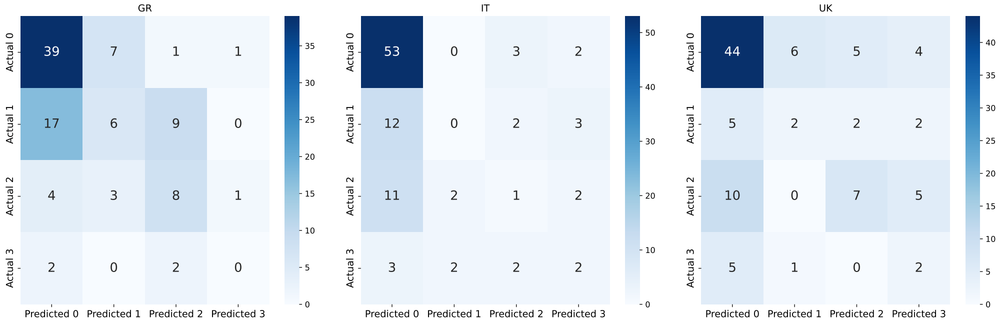


Figure B.8: Qwen2 LOOCV w/ Law confusion matrix.

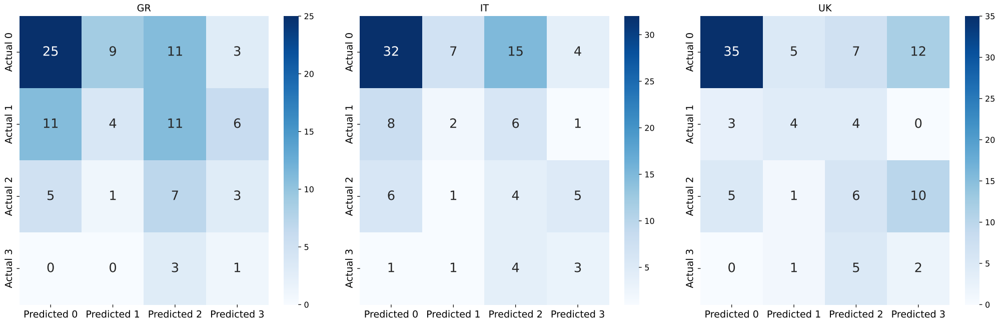


Figure B.9: Qwen2 4-shot w/o Law confusion matrix.

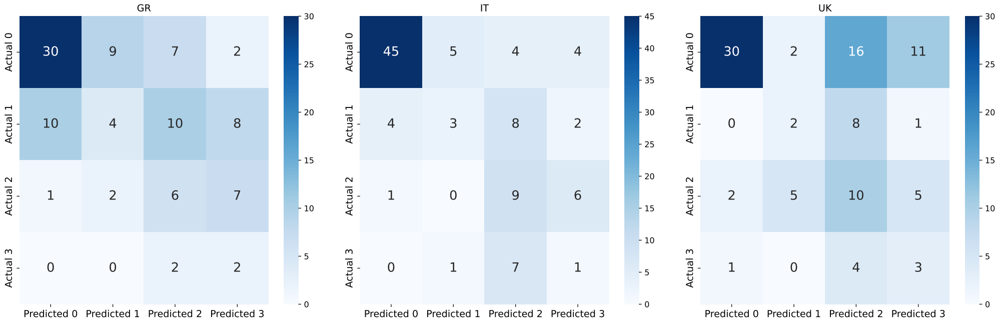


Figure B.10: Qwen2 4-shot w/ Law confusion matrix.

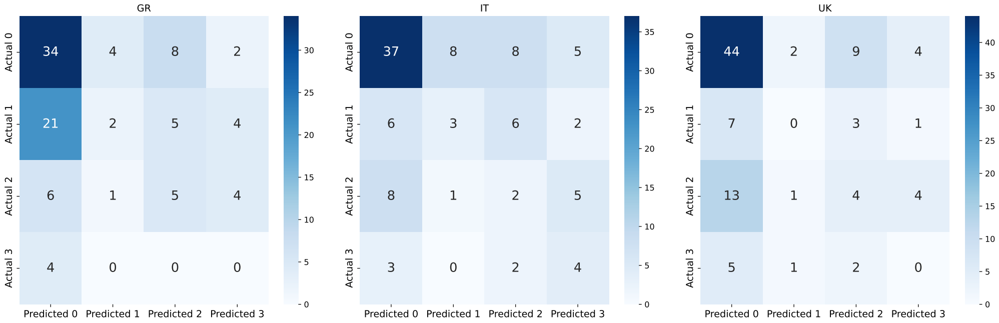


Figure B.11: Qwen2 8-shot w/o Law confusion matrix.

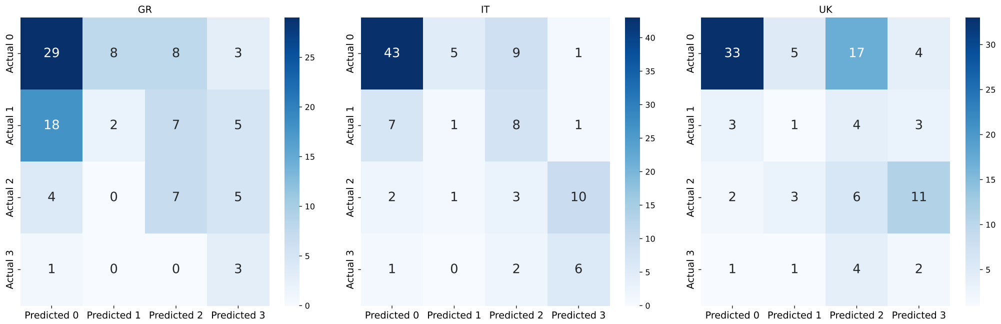


Figure B.12: Qwen2 8-shot w/ Law confusion matrix.

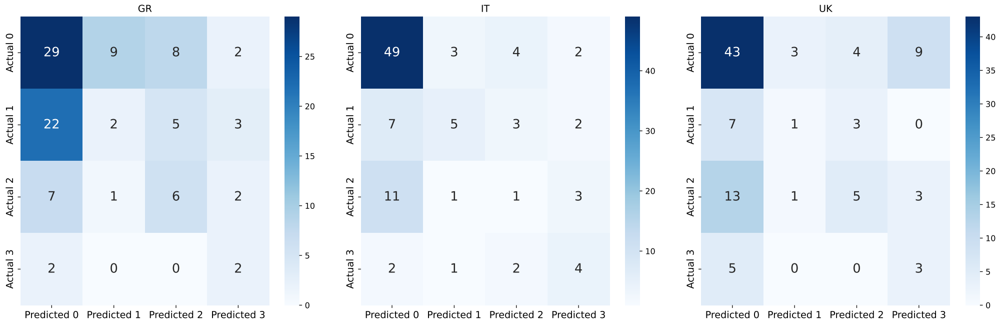


Figure B.13: Qwen2 12-shot w/o Law confusion matrix.

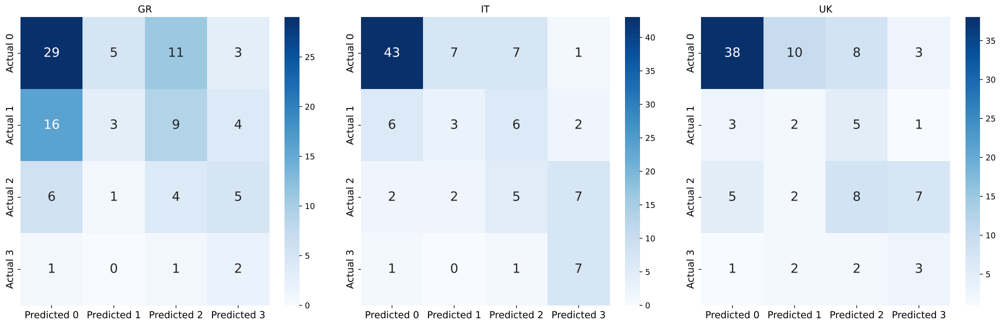


Figure B.14: Qwen2 12-shot w/ Law confusion matrix.

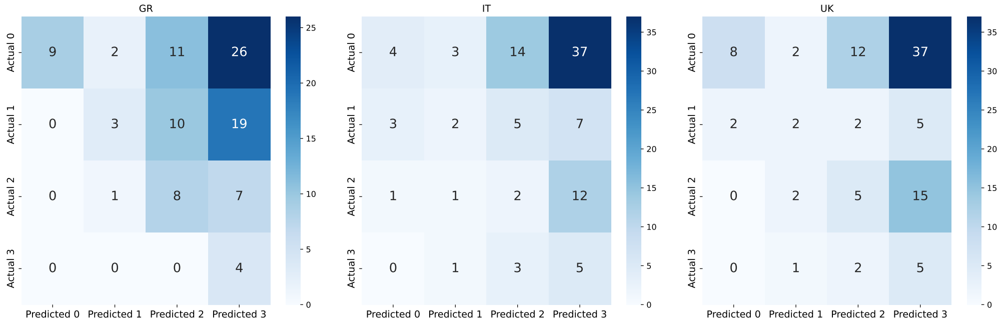


Figure B.15: Llama3 0-shot w/o Law confusion matrix.

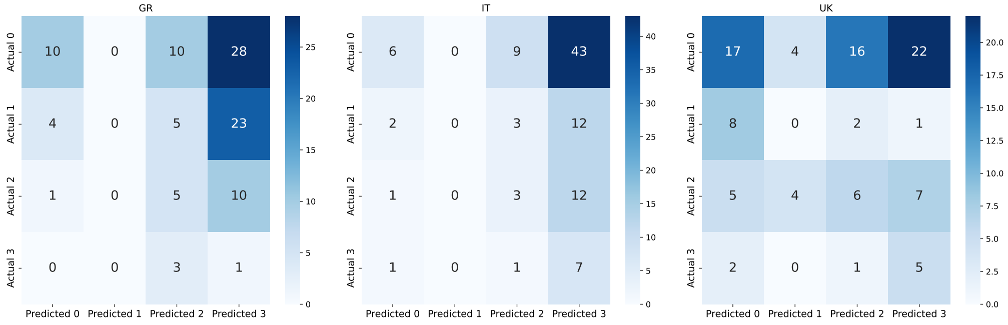


Figure B.16: Llama3 0-shot w/ Law confusion matrix.

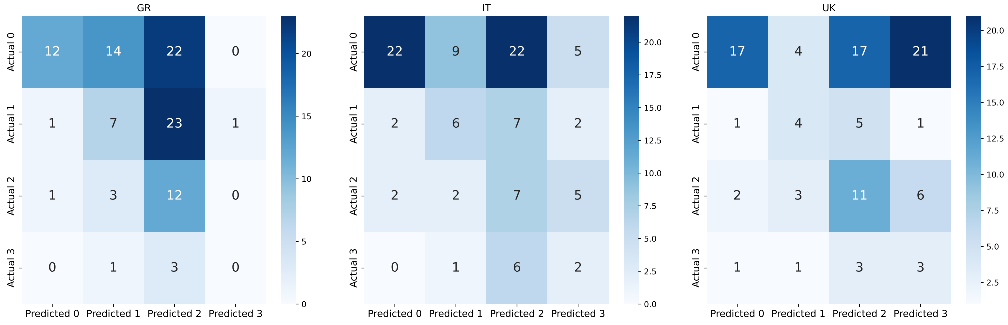


Figure B.17: Llama3 LOOCV w/o Law confusion matrix.

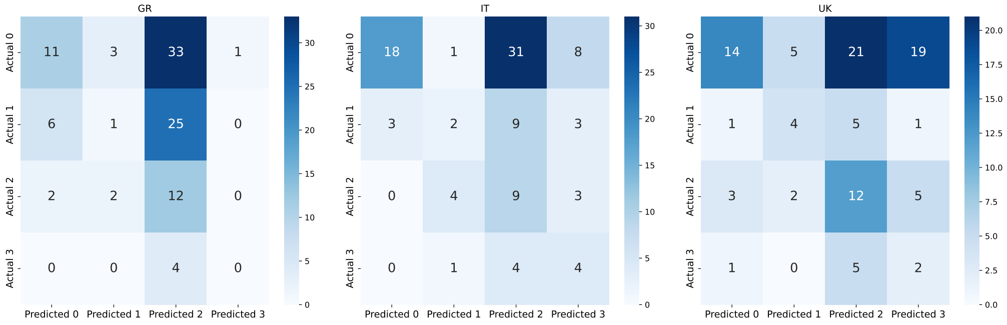


Figure B.18: Llama3 LOOCV w/ Law confusion matrix.

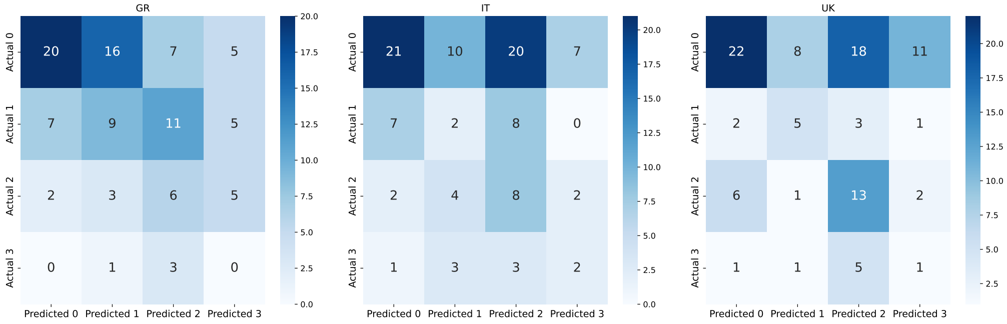


Figure B.19: Llama3 4-shot w/o Law confusion matrix.

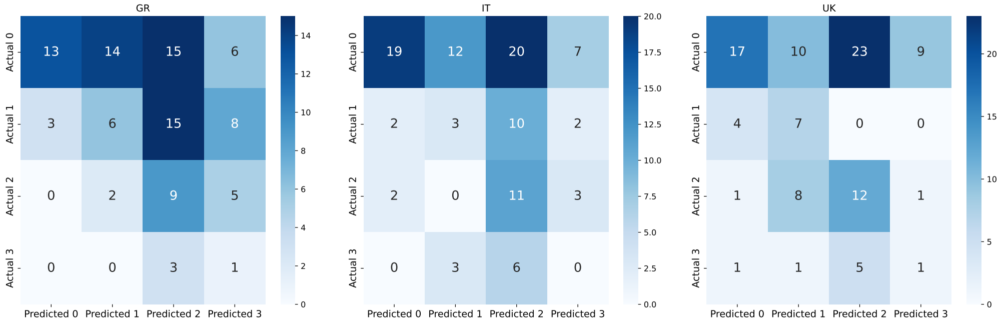


Figure B.20: Llama3 4-shot w/ Law confusion matrix.

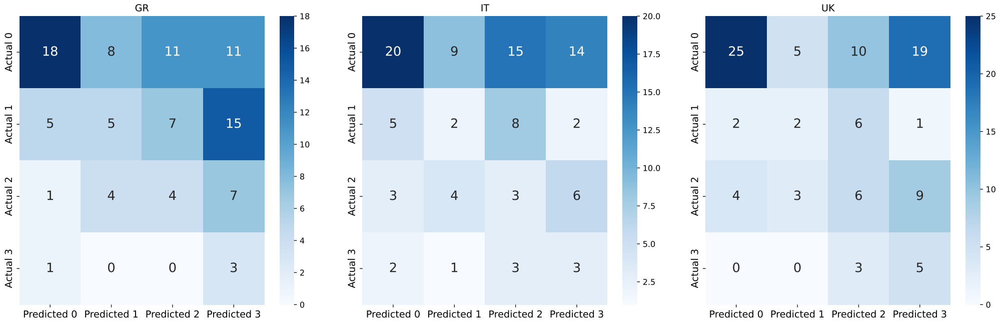


Figure B.21: Llama3 8-shot w/o Law confusion matrix.

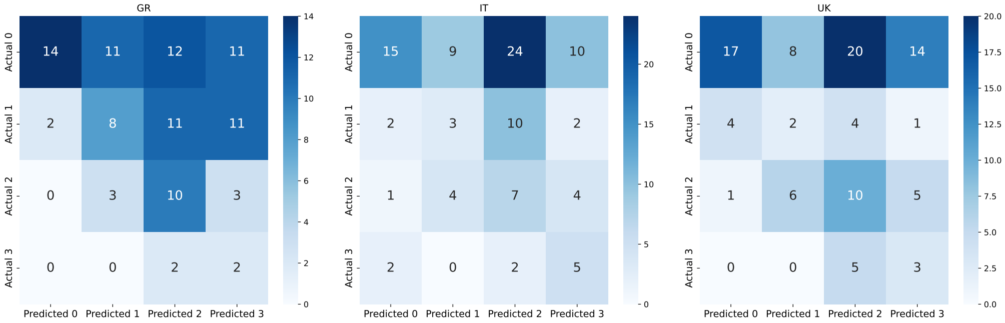


Figure B.22: Llama3 8-shot w/ Law confusion matrix.

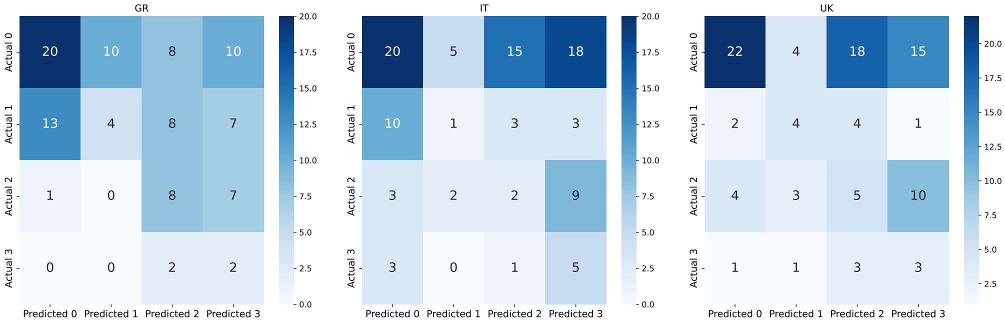


Figure B.23: Llama3 12-shot w/o Law confusion matrix.

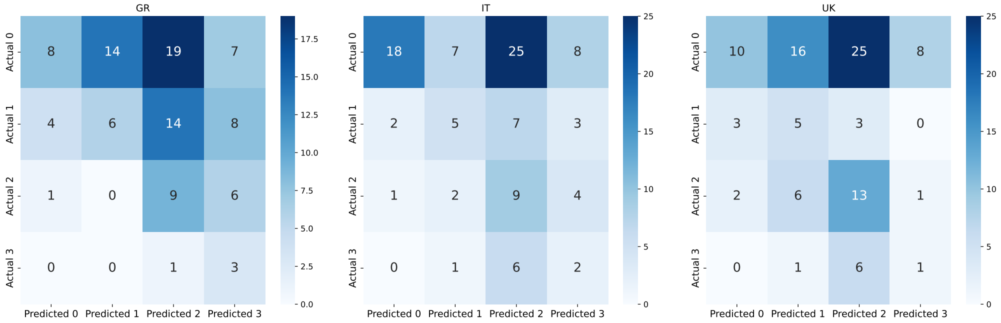


Figure B.24: Llama3 12-shot w/ Law confusion matrix.

Appendix C

More on Semantic Componential Analysis

C.1 Laws from the Global Handbook on Hate Speech Laws

We provide the full list of countries included and excluded from HateDefCon.

Included. Afghanistan, Albania, Algeria, Andorra, Angola, Argentina, Armenia, Australia, Austria, Azerbaijan, Belarus, Belgium, Bolivia, Bosnia and Herzegovina, Botswana, Brazil, Brunei Darussalam, Bulgaria, Cambodia, Cameroon, Canada, Central African Republic, Chad, Chile, China, Colombia, Croatia, Cuba, Cyprus, Czech Republic, Democratic Republic of Congo, Denmark, Djibouti, Estonia, Ethiopia, Fiji, Finland, France, Gabon, Georgia, Germany, Ghana, Greece, Guinea, Guinea-Bissau, Guyana, Haiti, Hungary, Iceland, India, Indonesia, Iran (Islamic Republic of), Iraq, Ireland, Italy, Japan, Jordan, Kenya, Kyrgystan, Latvia, Luxembourg, Myanmar, Malaysia, Malta, Mexico, Moldova, Monaco, Myanmar, Nepal, Netherlands, New Zealand, Norway, Oman, Pakistan, Sweden, Syrian Arab Republic, Tanzania, Timor-Leste, Trinidad and Tobago, Tunisia, Turkmenistan, Uganda, United Arab Emirates, United States of America, Uzbekistan, Venezuela, Zambia, Poland, Portugal, Romania, Russian Federation, Rwanda, Senegal, Serbia, Sierra Leone, Singapore, Slovakia, Somalia, South Africa, South Sudan, Spain, Sri Lanka, Switzerland, Tajikistan, Togo, Turkey, Ukraine, United Kingdom, Uruguay, Vietnam, Zimbabwe.

Excluded. Madagascar, Malawi, Maldives, Mali, Marshall Islands, Mauritania, Mauritius, Micronesia (Federated States of), Mongolia, Montenegro, Morocco, Mozambique, Namibia, Nauru, Nicaragua, Niger, Nigeria, North Macedonia, Palau, Panama, Papua New Guinea, Paraguay, Peru, Philippines, Qatar, Republic of Korea, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Samoa, San Marino, Sao Tome and Principe, Saudi Arabia, Seychelles, Slovenia, Solomon Islands, Suriname, Thailand, Tonga, Tuvalu, Vanuatu, Yemen.

C.2 Annotation Guidelines

Read and Understand the Definition. Carefully read the provided hate speech definition from the source material. Ensure you understand the context and content before proceeding with annotations.
Identify and Annotate Core Components. After reading the definition, review all available components listed in each Excel column. Annotate each component with 1 if the component exists in the definition. Annotate with 0 if the component does not exist in the definition. If a definition is very general, annotate with 1 on the column "General". If you see words/phrases like sexism, or sexist behavior, please annotate by putting one in gender.
Undefined Components. If you detect a component that has not been included in the columns, please use the column "Undefined Component" to write it down.
Add Comments. Feel free to add any relevant comments in the comments column.
Note: If a component or a morphological derivative of the component appears in the definition, mark the respective column with a positive annotation. For instance, if the predefined component is "abuse" and the component "abusive" is found in the definition, place a positive annotation (i.e., 1) in the "abuse" column.

Table C.1: HateDefCon annotation guidelines.

Table C.1 reports annotation guidelines for building HateDefCon via SCA.

C.3 Inter-annotator Agreement for Componential Annotation

Table C.2 reports pairwise IAA on HateDefCon.

Annotator Pair	**Average Kappa**
Annotator$_1$ vs. Annotator$_2$	0.75
Annotator$_1$ vs. Annotator$_3$	0.64
Annotator$_2$ vs. Annotator$_3$	0.64

Table C.2: Average Cohen's Kappa scores per annotator pair

C.4 Complete Component Hierarchy and Further Statistics

Table C.3 reports the fine-grained SCA hierarchy extracted from HateDefCon.

Framework	**Categories**
Target	**Demographics and Identity** Race/Ethnicity: Race, Ethnicity, Tribe, Color, Nationality, Regional Origin, Genetic Origin Nationality/Region: Nationality, Region, Place of Birth, Place of Origin, Place of Residence, Immigration Status Religion/Belief: Religion, Belief, Creed, Ideology, Philosophical Opinion, Philosophical Ideology, Worldview Gender/Sexual Orientation: Gender, Sexual Orientation, Gender Identity, Sex Change Disability: Disability, Physical Condition, Mental Capacity, Health Characteristics, Ability Age/Appearance: Age, Appearance, Generation
	Social and Economic Roles Occupation/Profession: Occupation, Profession, Job, Employment, Trade Union Membership, Calling Family Status: Family Status, Marital Status, Familial Status, Pregnancy Citizenship/Legal Status: Citizenship, Legal Status, Immigration, Veteran Status, Refugees, Nationality
	Social and Economic Class Socioeconomic Status: Economic Status, Social Class, Financial Status, Wealth, Poverty, Social Origin, Social Strata, Economic/Social Origin Caste/Tribe: Caste, Tribal Affiliation, Ancestry, Descent
Intent/Purpose	**Discrimination and Prejudice** Discrimination: Discrimination, Discriminatory Practices, Exclusion, Marginalization, Segregation, Denigration Prejudice: Prejudice, Bias, Stereotyping, Xenophobia, Ethnocentrism, Bigotry, Contempt, Superiority Humiliation: Humiliation, Demeaning, Belittling, Degrading, Ridiculing, Mocking, Stigmatizing, Inferiorizing, Dehumanizing
	Hostility and Aggression Violence: Physical Violence, Aggression, Abuse, Threat, Brutalization, Persecution, Terrorism, Hostility, Rancor Hate: Hatred, Ill-Will, Animosity, Abhorrence, Detestation, Malice, Anti-Semitism, Racism, Ethnocentricism Conflict: Conflict, Discord, Dissension, Sectarianism, Division, Social Unrest, Civil Unrest, War
	Social and Cultural Control Cultural Control: Cultural Manipulation, Propaganda, Ideological Imposition, Social Control, Social Hatred, Supremacy, Perpetuation of Norms, Customs, Traditions, Cultural Values Exclusion: Exclusion, Social Marginalization, Social Exclusion, Economic Exclusion, Stigmatization, Alienation, Isolation Suppression: Suppression, Silencing, Censorship, Restriction, Limitation of Rights, Deprivation, Harassment
Act/Means	**Verbal and Written Expressions** Insults: Insults, Pejoratives, Slurs, Offensive Language, Derogatory Language, Humiliation, Threat Defamation: Defamation, Slander, Vilification, Disparagement, Discrediting, Ridicule, Mockery Provocation: Provocation, Incitement, Inflammatory Speech, Sedition, Antagonism, Triggering, Threats Misinformation: Misinformation, Disinformation, Propaganda, Deception, Promoting Xenophobia, Racism, and Bigotry
	Physical Actions Violence: Physical Harm, Assault, Attack, Damage to Property, Brutalization, Persecution Exclusion: Exclusion, Segregation, Denial of Rights, Obstructing Rights, Deprivation, Harassment Cultural Actions: Desecration, Desecration of Symbols, National Flag Desecration, Denial of Cultural Identity
	Social and Cultural Manipulation Social Control: Manipulation of Social Norms, Cultural Domination, Supremacy, Cultural Stereotyping, Perpetuation of Prejudice Cultural Exclusion: Cultural Alienation, Cultural Stigmatization, Exclusion from Cultural Activities, Denial of Cultural Identity Economic Suppression: Economic Marginalization, Social Segregation, Restriction of Economic Opportunities, Denial of Economic Rights

Table C.3: Hate Speech Framework Hierarchical Structure.

C.5 Hate Speech Definitions for GHC

Table C.4 reports the definitions used in our experiments with LLMs on GHC.

No.	Definition
D_{ghc}	Language that intends to — through rhetorical devices and contextual references — attack the dignity of a group of people, either through an incitement to violence, encouragement of the incitement to violence, or the incitement to hatred.
D_{wiki}	Hate speech — a term that denotes speech intended to degrade, disturb, or cause violence or actions based on prejudice against persons or groups of people on the basis of their race, gender, age, ethnicity, nationality, religion, sexual orientation, gender identity, disability, language ability, moral or political views, socioeconomic class, occupation or appearance (such as height, weight, and hair color), mental capacity, and any other characteristic. The term refers to both written and oral communication, as well as some forms of behavior in a public place. Hate speech operates outside the law, speech that offends a particular person or group in terms of discrimination against that person or group. According to the law, hate speech is any speech, gesture or behavior, written text, or display that is prohibited because it is likely to incite violence or prejudice against or by a protected individual or group, or that degrades or intimidates a particular individual or group. The law may recognize the protected individual or group according to certain characteristics.
D_{dict}	Speech expressing hatred of a particular group of people.

Table C.4: Definitions of Hate Speech. D_{ghc} is the one used to annotate GHC, D_{wiki} is from Wikipedia, and D_{dict} is from the Meriam Webster Dictionary.

C.6 Definitions for Cross-Cultural Analysis

We select three definitions from our corpus. Table C.5 shows the prompt and the definitions used for the experiments.

Prompt

Read carefully the definition of ’hate speech’ provided. Your task is to classify the input text as containing hate speech or not. You can only rely on the definition provided. Respond only with YES or NO.
Definition: {definition}
Text: {text}
Answer:

Definitions
$\texttt{D}_{lev}$. Hate speech (HS) is formally defined as “any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic” Mulki et al. (2019)
$\texttt{D}_{isr}$. In this Article: “racism” – persecution, humiliation, degradation, a display of enmity, hostility or violence, or causing violence against a public or parts of the population, all because of their colour, racial affiliation or national ethnic origin. Publication of racist incitement is prohibited Article 144B: (a) If a person publishes anything in order to incite to racism, then he is liable to five years imprisonment. (b) For the purposes of this section, it does not matter whether the publication did cause racism, and whether or not it is true. Article 144C: Permissible publication (a) Publication of a true and fair report of an act said in section 144B shall not be deemed an offense under that section, on condition that it was not intended to cause racism. (b) Publication of quotes from religious scriptures or prayer books or the observance of a religious ritual shall not be deemed an offence under section 144B, on condition that it was not intended to cause racism. Article 144D: Possession of racist publication If a person holds a publication prohibited under section 144B for distribution, in order to cause racism, then he is liable to one-year imprisonment, and the publication shall be confiscated.
$\texttt{D}_{syr}$. Criminal Code Article 306: Any act, piece of writing or speech that is intended to or results in stirring sectarian or racial strife or inciting conflict between sects and the various elements of the nation shall be punished by imprisonment of six months to two years and a fine of one hundred to two hundred Syrian pounds, as well as with prohibition from exercising the rights mentioned in the second and fourth paragraphs of Article 65. Article 65: Every person sentenced to imprisonment or house arrest in misdemeanor cases is deprived throughout the execution of his sentence from exercising the following civil rights: A: The right to assume public employment and services. B: The right to assume jobs and services in managing the affairs of the civil sect or managing the union to which he belongs. C: The right to be a voter or elected in all state councils. D: The right to be a voter or elected in all sects and trade union organizations. E: The right to wear Syrian or foreign medals.
$\texttt{D}_{jor}$. Criminal Code Section 5: Crimes Harming National Unity and the Coexistence between the Nation’s Elements Article 150: Any writing or speech aims at or results in stirring sectarian or racial prejudices or the incitement of conflict between different sects or the nation’s elements, such act shall be punished by imprisonment for no less than six months and no more than three years and a fine not to exceed five hundred dinars (JD500). Audiovisual Media Law Article 20(l)(2) prohibits licensed broadcasters from broadcasting hateful, terrorist, violent or seditious material or from promoting religious, sectarian or ethnic strife.

Table C.5: Prompt and definitions used in the cross-cultural experiments.