Re-thinking Large Scale Hate Speech Identification: Beyond Common NLP Conventions and Supervised Machine Learning

Paula Cristina Teixeira Fortuna

TESI DOCTORAL UPF / year 2022

THESIS SUPERVISOR

Leo Wanner
co-supervisor: Juan Soler-Company

Department of Information and Communications Technologies

upf. Universitat Pompeu Fabra *Barcelona*

To my grandmother Adelaide, who, despite never having had the chance to learn to read or write, made sure that all of her children got the opportunity to go to school. This thesis would most likely not be possible without her initiative.

Acknowledgments

I am grateful to my colleagues, instructors, and friends at Pompeu Fabra University, ETIC for their support along this journey: Leo Wanner and Joan Soler-Company, respectively supervisor and co-supervisor, for their involvement and guidance throughout this project; Ahmed AbuRa'ed, Diana Ramírez-Cifuentes, Laura Pérez-Mayos, Marzieh Karimi-Haghighi, Mónica Domínguez and Roberto Carlini, for their friendship and the stimulating conversations! Alexander Shvets for his assistance during our collaboration. Lydia Garcia, not only for her administrative support, but also for her warm words whenever we need to communicate.

I couldn't forget to thank the people I worked with on the Stop PropagHate project and continue to collaborate with on subsequent scientific works: Sérgio Nunes for always being a source of positive feedback and words about my ideas and scientific contribution, Vanessa Cortez for her friendship, and Luís Braga Cruz for assisting me in evolving as a mentor myself during his Master thesis co-supervision.

I am grateful to the following researchers for their friendship, technical assistance, or both: João Rocha da Silva, Zeerak Talat, Wenjie Yin and Miguel Ramalho.

I couldn't forget the help I received from friends who heard my complaints about PhD life: Daniela Cardoso, Isabel Pratinha, João Maia, Luís Ribeiro, Miguel Sandim, and Sofia Reis. In academia, mental health can suffer, and it is not shameful to seek support, not just from friends, but also from professional therapists, for whom I am grateful, too.

During the last period of my PhD, the Women Climbing Porto community was especially vital in offering a supportive environment where I could disengage from intellectual abstract work, be mindful and recharge energies. Thank you!

And last but not least, I am thankful for the support of my family. Thanks to my mother for the values she passed me: no doubt patience, preserverance and hardwork are required in research and throughout a PhD. And, thanks to Ilaria Bonavita for her encouragement, love and infinite conversations and criticisms on Artificial Intelligence subjects.

During her PhD, Paula Fortuna is supported by the research grant SFRH/BD/143623/2019, provided by the Portuguese national funding agency for science, research and technology, Fundação para a Ciência e a Tecnologia (FCT), within the scope of Operational Program *Human Capital* (POCH), supported by the European Social Fund and by national funds from MCTES.

Abstract

The detection of hate speech in online spaces is traditionally conceptualized as a classification task that uses Machine Learning (ML)-driven Natural Language Processing (NLP) techniques. In accordance with this conceptualization, the hate speech detection task relies upon common conventions and practices in Artificial Intelligence, ML and NLP – among them interpretation of the inter-annotator agreement as a way to measure dataset quality and the use of standard metrics such as precision, recall or accuracy and benchmarks to assess model performance. However, hate speech is a highly subjective and context-dependent notion that eludes such static and disembodied practices. Their application results in definitorial challenges and the failure of the models to generalize across different datasets, two problems that I analyse in empirical studies. Furthermore, I critically reflect on the followed methodologies. I argue that many conventions in NLP are poorly suited for the problem and suggest to develop methods that are more appropriate for fighting online hate speech.

Keywords: hate speech detection, machine learning conventions, algorithmic challenges

Resum

Abordar el discurs de l’odi als espais en línia s’ha conceptualitzat com una tasca de classificació que utilitza tècniques d’intelligència artificial (IA), aprenentatge automàtic (ML) o processament del llenguatge natural (PNL). Mitjançant aquesta conceptualització, la tasca de detecció del discurs d’odi s’ha basat en les convencions i pràctiques comunes d’aquests camps. Per exemple, l’acord entre anotadors es conceptualitza com una manera de mesurar la qualitat del conjunt de dades i s’utilitzen determinades mètriques i punts de referència per inferir el rendiment del model. Tanmateix, el discurs de l’odi és un concepte profundament complex i situat que eludeix aquestes pràctiques estàtiques i incorpònies. En aquesta tesi aprofundeixo en els reptes de definició i les dificultats pel que fa a la generalitzaci´o de models, dos problemes que analitzo amb estudis empírics. A més, reflexiono críticament sobre les metodologies seguides, argumento que moltes convencions en PNL són poc adequades per al problema i animo els investigadors a desenvolupar mètodes més adequats per combatre el discurs d’odi en línia.

Keywords: detecció de discurs d’odi, convencions d’aprenentatge automàtic, reptes algorísmics

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

Online hate speech and, in particular, hate speech in social media, is a cause for growing concern. For instance, in an evaluation of 263 million conversations in the United Kingdom and United States between 2019 and mid-2021, 50.1 million debates about, or instances of, racial hate speech have been discovered (rep, 2022).

Social media companies such as Facebook, Twitter, Instagram, etc. are legally obliged to monitor their platforms and eliminate timely hate speech. In an attempt to comply with these obligations, they invest a lot of human resources to identify hate speech. But human moderation is expensive and, in addition, often implies a significant negative psychological impact on the moderators (Roberts, 2019). This calls for automation of the procedure. Subsequently, the automatic detection of hate speech has become a popular research topic in Natural Language Processing (NLP), with a substantial number of papers at leading conferences (e.g., the main conference and the Findings of Association for Computational Linguistics (ACL) 2021 featured nine papers with titles related to hate speech, and the Conference of Empirical Methods in Natural Language Processing (EMNLP) / Findings of EMNLP 2021 five papers), targeted workshops (such as, e.g., the series of the Workshops on Online Abuse and Harms) and shared tasks (such as, e.g., International Workshop on Semantic Evaluation (SemEval) 2021 Task 5: Toxic Spans Detection) dedicated to it.

Classifiers (or, in more generic terms, supervised machine learning, ML) based on Artificial Intelligence (AI) and NLP are considered by the state of the art as most appropriate for hate speech identification: trained on material annotated in terms of a restricted number of hate speech categories they are supposed to be able to correctly assign a given text sample to one of those categories or to the 'not-hate speech' category.

As the field of NLP has evolved in the past decades, common conventions have been established and proved essential in the context of supervised ML. These conventions correspond to practices put often into action to ensure the quality of research in the area. They include, among others, the necessity of comprehensive dataset annotation guidelines and the conceptualization of the inter-annotator agreement as a way to measure dataset quality. In addition, specific metrics and benchmarks are used to assure models generalize to new data and contexts. However, hate speech is a deeply complex and situated concept, and distinct from other linguistic phenomena which have been addressed by supervised ML. Therefore, it is legitimate to ask whether we can assume that these conventions are valid for a complex problem such as hate speech. And most importantly, what are the limitations of the traditional supervised ML methodology in the specific case of hate speech?

In this thesis, conventions established for supervised ML are under scrutiny since the answer to the question on their suitability for a specific text genre such as hate speech has far-reaching consequences: if these conventions are found to be unsuitable, the application of classification algorithms to the genre in question must be reconsidered.

1.1 Motivation of the Thesis

Machine learning-based classification techniques have shown to be a successful instrument for the identification of documents with some specific characteristics, in the simplest case by a binary division of a given document collection into documents that possess these characteristics and those that do not. Therefore, classifiers are also considered in the state of the art to be the solution to the detection of online hate speech.

Over the last decade, a considerable body of work has been carried out in the area of automatic identification and classification of hate speech and related phenomena in terms of fine-grained categories such as ‘racism’, ‘sexism’, ‘hate speech’, etc.; and general categories such as ‘offensive’, ‘abusive’ or ‘toxic’ text. Due to their large-scale data processing capabilities and promising benchmarking results, classifiers have been marketed as an effective solution to detect these categories. Gradually and sparingly, however, work on hate speech identification has been generating concerns. It has been demonstrated that published research raises significant challenges, among them, that: (a) in terms of data generation, hate speech collections are not random and rely on particular sampling strategies; (b) after training, models do not generalize to new data; and (c) as a consequence of the sampling strategies, models contain social bias, since the chance of identifying material as hateful depends on the group of the speaker. In cases when previous works on the limitations of classifiers for hate speech focused on data sampling, model generalization and bias, e.g., Vidgen and Derczynski (2021); Waseem et al. (2018); Davidson et al. (2019); Nozza (2021), they have been incomplete in their examination of the complexities of hate speech definitions and models when applied in new contexts. In addition, they neglected to investigate the importance of certain conventions for classification in the validation of previous research.

This thesis takes up these shortcomings. By analyzing standard classification conventions and investigating their application to hate speech, I highlight new open questions in the field with regard to the suitability of existing state-of-the-art models to this problem.

Showing concerns for the application of classifiers to the problem of hate speech does not exclude the use of algorithms or large-scale analyses for aiding to solve this problem. We need transparent algorithms for processing of vast amounts of data that can assist us in understanding some of the phenomena occurring in the internet, as well as which information is spread and how, in order for people and society to devise ways to deter them. Given the necessity for certain big data tools, this thesis sheds some light on the kind of solutions that may assist us in this endeavor.

1.2 Research Goals

This thesis has as main goal to investigate the limitations of the technology used to date to identify hate speech. It does so by analyzing the standard practices, namely in terms of how the task of hate speech detection is being formulated and by further investigating the inadequacy of some common conventions for classification.

When I apply the term "common conventions" to classification, I refer to the procedures generally accepted by the communities using ML and AI technologies as part of the pipeline to build classifiers. In this thesis, the conventions that I am analysing are: (i) computing inter annotator agreement, (ii) computing ground truth by relying usually on three annotators, (iii) computing benchmarks with evaluation metrics such as precision, recall, F1, and (iv) using train-test division for evaluating model generalization.

The goals of this thesis are:

1. Analyzing definitorial challenges when framing hate speech identification as a classification problem.

2. Analyzing conventions for classification when applied to the problem of hate speech detection.

3. Providing methods that support a better understanding of hate speech concepts when annotated in datasets.

4. Increasing hate speech dataset compatibility.

5. Finding new conventions for testing model generalization.

6. Understanding the lack of model generalizability for hate speech.

7. Investigating the continuation of application of classifiers to hate speech detection.

1.3 Structure of the Thesis

To investigate the problems when using classifiers for hate speech, I follow two main approaches: first, I acquire insight via critical literature review; second, to further study some topics, empirical research is undertaken. The chapters of the thesis are structured so as to:

- Start with an Introduction (see Chapter 1) on the thesis motivation (cf. Section 1.1), goals (see Section 1.2) and structure (see Section 1.3).

- Present the Background and fundamental knowledge (see Chapter 2).

- Consider commonly used classification conventions and investigate its adequacy to the problem of online hate speech (cf. Chapter 3). I will specifically focus on the modeling of hate speech within the standard methodology for classification, which involves several well-known steps: 1) the collection of data, 2) annotation of this data, 3) deployment of classification algorithm/s, and evaluation of results.

- Investigate the lack of definitional clarity for hate speech datasets and the importance of improving this aspect when building classifiers (see Chapter 4). I will answer to this demand by creating a conversion between classes annotated in the different datasets needed for the experiment presented in the following chapter.

- Apply new model evaluation procedures capable to give a better sense of the generalizability of models for hate speech classification (see Chapter 5).

- Discuss the required steps for continuing to explore supervised classification as a solution to the problem of online hate speech, as well as why the use of classifiers is incompatible with an anti-colonial and systemic approach to combating hate speech (see Chapter 6).

- Finish with a Conclusion (see Chapter 7) on the thesis contributions, limitations and future works.

Chapter 2

BACKGROUND

2.1 Background Theories and Definitions

This thesis investigates the difficulties encountered when using classifiers and the conventions underlying them to detect online hate speech. It is thus critical to analyze the many dimensions and development of hate speech-related notions, as well as the theories upon which our study is based. I begin the foundations part by reflecting on the definition of hate speech.

2.1.1 Defining Hate Speech

Over the last decade, algorithms for identifying hate speech have been created relying on a range of different concepts. Defining hate speech and applying this definition to annotated data is not easy, and much work has been spent on this conceptualization task. Several surveys sought to define and contextualize the topic, including Schmidt and Wiegand (2017); Fortuna and Nunes (2018); Poletto et al. (2021). As a result, a typical hate speech definition would be succinct and centered on general subset of the minorities who are targeted by hate speech and on how violence is produced against them; consider, for instance:

> “Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humour is used.” (Fortuna and Nunes, 2018).

The definitions of hate speech, which are often used in the field, e.g., including Schmidt and Wiegand (2017); Fortuna and Nunes (2018); Poletto et al. (2021), will prove inappropriate (see Chapter 3), underlining the need to seek for wider concepts. One example of these is a systemic discrimination approach to define and identifying hate speech which is followed in this thesis. According to a systemic discrimination approach, hate speech is speech that reflects and maintains systemic discrimination against the group to which the target is thought to belong, therefore inhibiting targets from behaving freely (Gelber, 2021). A systemic approach to hate speech does not rely on the presence or detection of hate and it does not rely on the use of specific vocabulary. Instead, it recognizes that not all group identities face systemic discrimination, it recognizes a speaker’s capacity to harm in a systemic manner, and it suggests a non-punitive, discourse-based strategy for responding to hate speech (Gelber, 2021).

Before diving into the specifics of systemic discrimination notions, I will compare and contrast various types of online verbal abuse. In the domains of AI and ML, it is crucial to examine such concepts since they have been related to hate speech.

2.1.2 Hate Speech and Other Online Abuse

Hate speech is a unique phenomenon in terms of its definition, legal implications, and societal effects. However, in the ML and AI research, it has been studied repeatedly in conjunction with other terms, which I will list here, in order to clearly set the boundaries between the diverse phenomena. I have included a glossary of terms based on past surveys (Fortuna and Nunes, 2018; Poletto et al., 2021), as well as dictionary definitions.

Hate It is an unjustified display of hostility (Tarasova, 2016), which is often used interchangeably with the term hate speech, although the second focuses on group discrimination rather than the broad expression of a hateful feeling.

Cyberbullying Aggressive and deliberate act committed repeatedly and over time by a group or person utilizing electronic forms of communication against a victim who is unable to protect himself or herself (Chen et al., 2012). Bullying another person may take on a variety of forms, and hate speech can be utilized to accomplish this objective. Even when used to interpersonal bullying, however, it involves the replication of group dynamics and the exercise of power, violence, or silence by a dominating group against a minority.

Flaming Is a term that refers to aggressive, vulgar, and threatening remarks that may disrupt community engagement (Guermazi et al., 2007). Whereas flaming is broad and directed at a particular participant in the context of a discussion, hate speech is a verbal form of prejudice that may occur in any setting.

Extremism Ideology connected with radicals or hate organizations that promotes violence, often with the goal of segmenting people and regaining status, and in which outgroups are portrayed as both offenders and inferior populations (McNamee et al., 2010). Extremist discourses often make use of hate speech. However, these discourses address a variety of additional issues (McNamee et al., 2010), including recruitment of new members, government and social media demonization of the in-group, and persuasion (Prentice et al., 2011).

Radicalization Online radicalization is analogous to the notion of extremism and has been researched across a variety of subjects and contexts, including terrorism, racist groups, and nationalism (Agarwal and Sureka, 2015). Extremist discourses, like radicalism, may use hate speech. However, radical discourse often includes references to war, religion, and unpleasant emotions (Agarwal and Sureka, 2015), while hate speech may be more subtle and based on stereotypes.

Profanity Language that is blasphemous or vulgar. A swear word, a spoken oath (lin, 2021b). Hate speech may or may not utilize such language.

Obscenity A phrase or statement that is very unpleasant or sexually disturbing (lin, 2021a). These phrases may or may not be used in hate speech.

Threat A declaration of purpose to cause someone pain, injury, property damage, or other hostile action in retaliation for something done or not done (lin, 2021b). Such declarations may or may not constitute hate speech.

Insult A sarcastic or contemptuous statement or behavior lin (2021a). Hate speech may include any kind insults, but it can also occur without mentioning any of those.

Additionally, certain umbrella words have been used often in the area of NLP to refer to online abuse and harms such as toxic, abusive, offensive, and aggressive language.

Toxicity Toxic remarks are those that are harsh, insulting, or unreasonable and are likely to cause a person to withdraw from a conversation (Jigsaw, 2019b). The category of toxicity is subdivided into threat, severe toxic, insult, obscene, and identity hate, the last of which corresponds to hate speech. Hate speech is distinct from toxicity since it may not always cause individuals to withdraw from discussions.

Abusive language The term "abusive language" was used to refer to hurtful (Nobata et al., 2016) or rude language, as well as violence used to be cruel to someone (lin, 2021a). This language may or may not qualify as hate speech.

Offensive language Offensive language is language that causes someone to feel resentful, unhappy, or irritated lin (2021b), or is likely to make others angry or upset lin (2021a). An offense may or may not qualify as hate speech.

Aggressive language Aggressive language is language that attacks, confronts (lin, 2021b), or it is angry and violent toward another person (lin, 2021a). Aggressive language may or may not qualify as hate speech.

Racism, Sexism or any type of prejudice Racism and sexism (or other class-based prejudice terms) refer to the idea that some individuals are inferior than another because of race or gender identity. As a result, we believe that racist or sexist discourses are hate speech forms.

2.1.3 Hate Speech as Verbal Discrimination

By understanding terms such as "social categorization", "discrimination", "stereotypes" and "prejudice", it is easier to comprehend hate speech. They are defined here since those concepts are also utilized in this thesis. Besides, the terms "stereotype", "prejudice" or "discrimination" are part of the everyday vocabulary raising the risk of vagueness and misinterpretation. To prevent an oversimplification of such constructs, I define them throughout this section, drawing extensively on social psychology literature.

Discrimination Is defined as unjustifiable negative behaviour aimed against a group or its members, including both actions directed at and judgements or decisions made about group members (Al Ramiah et al., 2010). While discrimination is often linked with individual behavior, it may also be extensively enforced via institutional structures and laws. Thus, processes at the individual, institutional, and cultural levels may interact to confer structural privileges on certain groups and/or impose penalties on others. Biases at the individual level of analysis indicate higher-order institutional and cultural biases due to the interrelated nature of these multilevel processes. These effects are often hidden by explanations or beliefs, enabling discrimination to go undetected and unacknowledged (Baumeister and Finkel, 2019). While discrimination is a kind of behavior, it is associated with and underlies certain cognitive processes, which I will discuss in more detail in the following paragraphs in order to get a better understanding of the cognitive components of hate speech, as well.

Social categorization Is the cognitive process by which individuals are categorized into groups, and a critical component of social functioning (Kawakami et al., 2017). According to (Tajfel et al., 1971), no intergroup contact can occur without a separation of the social environment into "us" and "them", or ingroup and outgroup. This simplification happens because concentrating on individuals requires much more cognitive effort and resources (Fiske, 2012). Once assigned to a group, a person is regarded to be similar to the other members of that group, and we attribute to her the traits that we believe all members of that group possess, which eliminates any ambiguity about future conduct of that person. As a result of social categorization, two biases develop: preference for the ingroup and avoidance or hostility toward the outgroup (Brewer et al., 1999).

Stereotypes Baumeister and Finkel (2019) define stereotypes as cognitive schemas summarizing the characteristics linked with specific groups and the group's anticipated social roles. Stereotypes can be verbalized through hate speech as some dataset annotation procedures identified (Sanguinetti et al., 2020). Although stereotypes are portrayed as a negative flawed thinking process, comparable to categorization and social categorization theory, research has concentrated on stereotypes' useful features in simplifying a complex environment. They are now considered to be cognitive schemas, often rooted in culturally held beliefs used to process information about others. While stereotypes typically include evaluative meaning and are often associated with prejudice, they do not have to be unfavorable.

These cognitive processes do not imply negative intergroup interactions in every instance. Nonetheless, we cannot overlook the drawbacks of such simplifications, such as the emergence of negative biases such as in the case of prejudice.

Prejudice Is a personal attitude toward groups and their members in order to maintain social hierarchy (Dovidio et al., 2010). For example, prejudice against women may have a "hostile" component that reflects a negative attitude toward women who stray from a conventional subservient role, as well as a "benevolent" component that supports women's subordinate position (Glick and Fiske, 2001). Thus, prejudice does not have to be associated with negative views against a target group; it may also be associated with apparently favorable attitudes toward an outgroup that result in discriminatory action (Baumeister and Finkel, 2019). This distinction is critical for the issue of identifying hate speech, since positive utterances about a group may underline discriminatory ideology as well.

Although the above terms are often used interchangeably in common speech, we should differentiate discrimination from prejudice and stereotyping so that we can get a more detailed image of how they relate with the production of hate speech. While discrimination is defined as *behaviors* directed toward a particular group of people, stereotypes are oversimplified *generalizations*, and prejudice is defined as the *emotions and attitudes* an individual has toward certain categories. Given that hate speech is one among other possible behaviors to discriminate and express prejudiced attitudes or stereotypes, intervening on hate speech production and spreading (behavior level) without aiming to change underlying prejudice and stereotypes (cognition level) may become ineffective in reducing actual discrimination. For this reason, in this thesis, I support approaches to fight online hate speech that allow intervention to act not only at the behavioral level but also at the cognitive one.

In the next section, I expose notions of privilege, marginalization and colonialism. These are necessary to comprehend how the existing limits of hate speech detection systems threaten minorities (see Chapter 6).

2.1.4 Reinforcing Privilege and Marginalization with Algorithms

In this thesis, I will discuss limitations of AI and ML systems for detecting hate speech. When defining hate speech, there is not an objective concept and such definition requires someone to decide who can be the target of hate speech and which comments will be considered an aggression toward that target. Such definition confers the privilege to control the discourses and decide which utterances can be reproduced or punished.

Privilege

When we debate prejudice, we are taught to define it as direct and specific acts of cruelty perpetrated by members of one group against members of another. This is not, however, an accurate depiction of the whole scenario. McIntosh (1988) discusses how we have been socialized to overlook a subsidiary element of discrimination: privilege. She focuses on privilege in particular as an unseen bundle of unearned advantages designed to remain oblivious. Privilege operates in online platforms, when defining hate speech, and through applying classifiers as well: certain groups create, design, define, comprehend, or utilize those tools, while others are subjected to their application without controlling the limits and conditions of its usage (see Chapter 6.2).

Algorithms as Weapons of Colonization

Colonization is a term associated with past periods in human history, however the processes of colonization still take place in the performance of the colonial present (Gregory, 2004). The word colonization connotes a relationship of structural dominance aimed at preserving resources for dominant objectives (Mohanty, 1988). For instance, hate speech has a role in colonial dynamics: the dominant cultures attempt to silence the other and retain their own power and privileges. Typically, dominant cultures portray others as exotic, bizarre, or alien, and this stereotyping of the other is often expressed and conveyed via hate speech.

Recent research has been discussing how AI and ML are instruments capable of enabling and sustaining colonialism and are likewise reliant and reliable tools on colonial dynamics (Couldry and Mejias, 2021). Decolonial movements and theories arise to examine both past and current manifestations of colonialism, and this also when applying digital technologies. In one of these efforts, according to Ricaurte (2019), a decolonial lens should illuminate how power imbalances manifest as digital colonialism and algorithmic violence. Data-extractivism is a term applied in this context[1]. It draws a parallel between information management and mining by describing data as a raw material that may be mined, and converted into other commodities with added value. Consumers appear as a natural supply of raw materials for research and technological development without control or fair compensation for the wealth they assist to generate. An example of data colonialism is portrayed by Noble (2018) in her book "Algorithms of Oppression". The author discusses how search engines behave differently and stereotypically when users enter keywords such as "black girls" vs. "white girls" (this by 2018). The engine's output would provide a hypersexualized and frightening picture of black femininity in contemporary culture. Noble (2018) argues the existence of a biased set of algorithms that favor whiteness and discriminate against people of color, particularly women of color. Data colonization concepts

[1] Extracted from `http://imaginacionmaquinica.cl/data-extractivism`

apply also to the decolonization of hate speech classifiers which I will discuss in a later chapter (see Chapter 6).

2.2 Algorithms Used in the Thesis

This section discusses the most relevant algorithms and data processing techniques used throughout this work.

2.2.1 Used Pre-processing Techniques

One must prepare high-quality data by pre-processing the raw information after data collection. One common approach is to remove certain sentence elements from text and social media data.

Stop Words Removal Stop words appear often in documents but have little significance since they are employed to connect words in a phrase. It highly depends on the purpose of the application what is considered to be a stop word and it is widely accepted that stop words do not add context or meaning to textual writings. As a result, one method is to eliminate them (Kannan et al., 2014).

Data Cleaning Common procedures when cleaning social media data is to remove from text IPs, hashtags and user-names.

Another often used strategy is to normalize data such that the available values fall within specified ranges.

Z-Score Normalization When the minimum and maximum values of attribute A are unknown, z-score normalization changes the values of a feature based on the mean and standard deviation of the feature (Al Shalabi et al., 2006). This normalization follows the expression:

$Z = \frac{v-\mu}{\sigma}$

In the expression, v refers to the original value, μ is the mean value of the feature and σ is the standard deviation of the feature.

2.2.2 Converting Text to Model Input

Depending on the experiment, I employ different methods in this thesis to convert unstructured text data to model input.

Bag of Words (BOW) Each word in the document is considered a feature. When single words are employed as features as in BOW, the resulting technique is referred to as a unigram word model. (Pustejovsky and Stubbs, 2012)

Word Embeddings The term word embedding refers to the representation of words, often in the form of a real-valued vector that encodes the meaning of the word in such a way that adjacent words in the vector space are predicted to have comparable meanings. Word embeddings may be produced in a variety of methods by mapping words or sentences from the lexicon to vectors of real numbers. I employ fastText in this thesis, a tool for which the word-embeddings are a result of classification. The fastText algorithm solves uses sub-word n-gram information, which may be used to determine the order relationship between characters and more accurately capture the core meaning of words (Mikolov et al., 2018).

2.2.3 Used Supervised Machine Learning Methods

While processing and analyzing data it is possible to differentiate between supervised and unsupervised ML methods. Supervision refers to the process of accumulating information about a specific task via the use of examples. More precisely, supervised approaches are always aimed at developing a model that generalizes by applying an algorithm to a collection of known data points in order to obtain insight into an unknown set of data (Kühl et al., 2020; Goldberg, 2017).

In the scope of this thesis, I use a wide range of supervised classification algorithms.

Support-Vector Machine (SVM) The idea behind Support-Vector Machine (SVM) techniques is to map training examples to points in space in order to maximize the width of the gap between the dataset's existing categories. New instances are then mapped into that same space and projected to belong to one of the categories depending on which side of the gap they fall on. In more technical terms, a support-vector machine, creates a hyperplane or group of hyperplanes in a high- or infinite-dimensional space. Intuitively, a successful separation is obtained by the hyperplane with the greatest distance to the closest training-data point of any class, since the bigger the margin, the smaller the classifier's generalization error (Hastie et al., 2009).

Random Forest Random forest is an ensemble learning approach that works by building a large number of decision trees during training (Biau and Scornet, 2016). Typically, decision tree algorithms work top-down, picking a feature at each level that best separates the collection of cases in terms of objective class. There are several metrics for determining which division is the best, one of which is information gain. In general, such measures evaluate the homogeneity of the target class across data splits. When all of the data inside a data split belongs to the same class, tree algorithms ground to a stop. For classification problems, the random forest output is the class chosen by the majority of trees. This is a general-purpose classifier with weak statistical assumptions.

Bagging Bagging is a ML ensemble technique that has been developed to increase the stability and accuracy of classification algorithms commonly used to reduce variance within a noisy dataset (Bühlmann, 2012). Given a standard training set D of size n, bagging generates m new training sets, by sampling from D uniformly and with replacement. Independent models are then trained, and the majority of those predictions result in a more accurate classification.

Convolutional Neural Network (CNN) A CNN is a deep neural network that takes its name from the mathematical linear operation between matrices called convolution (the function resulting from the multiplication of functions). CNNs differ from ordinary Neural Networks in that the neurons in one layer do not necessarily link to all the neurons in the following layer, but to a subset of them. CNNs may feature convolutional, non-linear, pooling, or fully connected layers (Gu et al., 2018).

FastText FastText is a off-the-shelf classification model. It is similar to the CBOW model of Mikolov et al. (2013), a model that tries to predict the target word by trying to understand the context of the surrounding words. In this thesis I ran the model in its version 0.9.2 with 300 dimension-FastText pretrained vectors (Mikolov et al., 2018), Skipgram Hierarchical softmax loss function, learning rate of 1.0, considering 1 as minimal number of word occurrences, bi-grams, and 25 epochs.

BERT BERT Devlin et al. (2019) makes use of Transformer, an attention mechanism that learns contextual connections between words (or sub-words) in a text. Transformer has two independent systems - an encoder that reads the text input and a decoder that generates a prediction for the job. Since BERT's purpose is to construct a language model, just the encoder mechanism is used. BERT considers both left and right context of words in all layers and pre-trains deep bidirectional representations from unlabeled text. As a consequence, the pre-trained BERT model may be fine-tuned by adding a single extra output layer without requiring significant changes to the task-specific architecture. In this thesis I use $\text{BERT}_{\text{LARGE}}$ (L=24, H=1024, A=16) with 340M parameters in total, which outperformed $\text{BERT}_{\text{BASE}}$ across all tasks, especially those with very little training data (Devlin et al., 2019), as is the case with some of our datasets. We use a batch size of 32 and fine-tune for 3 epochs over the data of all datasets. The dropout probability is set to 0.1 for all layers; the Adam optimizer is used with a learning rate of 2e-5.

ALBERT Lan et al. (2020) is A Lite BERT, a well-known technique for unsupervised language representation learning. ALBERT employs parameter-reduction approaches that enable large-scale setups, circumvent earlier memory constraints, and improve model degradation behavior. We use ALBERT$_{\text{XXLARGE}}$ (L=12, H=4096, A=64) model with about 70% of BERT$_{\text{LARGE}}$'s parameters for an available trained ALBERT model Lan et al. (2020), which we fine-tune to all nine datasets as described in Lan et al. (2020). We use a batch size of 32, the dropout probability is set to 0.1 for all layers and the Adam optimizer is used with a learning rate of 1e-5.

Perspective API It is an API that given an input text provides several classifiers for toxicity and other different categories (among others, ’identity hate’). The classifier uses Convolutional Neural Networks (CNNs) trained with GloVe word embeddings fine-tuned during training on data from online sources such as Wikipedia and The New York Times. The Perspective API was applied as a baseline system for toxicity detection, without any additional training on a contest data and obtained a very competitive result (12th out of 103 submissions, F1 of 0.79) indicating the possibility for this model to be used with new samples with good performance Pavlopoulos et al. (2019).

2.2.4 Used Procedures for Algorithm Evaluation

To evaluate any classification system it is common to start by building a confusion matrix (Jurafsky and Martin, 2009).

Confusion matrix A confusion matrix is a two-dimensional table that depicts how an algorithm performs with respect to the human gold labels, using two dimensions: system output and gold labels. True positives and negatives are documents that are accurately identified while false positives and negatives are incorrectly labeled instances (see Table 2.1).

Table 2.1: Confusion matrix

system output / gold labels	**HS**	**Not HS**
HS	True Positive (TP)	False Positive (FP)
Not HS	False Negative (FN)	True Negative (TN)

Accuracy Inquires what percentage of all observations the system accurately classified. While accuracy may seem to be a natural statistic, we seldom utilize it for text categorization tasks. This is because accuracy suffers when classes are uneven and precision and recall are used instead. Accuracy is given by:

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

Precision Precision is defined as the proportion of relevant occurrences within the positive instances retrieved. It is defined as:

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

Recall Recall quantifies the proportion of items in the input that were properly classified as positive by the system. It is defined as:

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

Thus, unlike accuracy, precision and recall place an emphasis on true positives. There are several methods to define a measure that combines elements of both. The F1-score is the simplest of these combinations.

F1-score The F1 score is the harmonic mean of the precision and recall and less affected than accuracy by the proportion of classes.

$$F1score = 2 * \frac{precision*recall}{precision+recall}$$

2.2.5 Feature Importance

Feature importance techniques have as goal to understand the features with greater importance for the classification.

Permutation feature importance algorithm The permutation feature importance is defined as the drop in a model's score caused by randomly shuffling a single feature value (Altmann et al., 2010). Permutation feature importance is a model inspection approach that may be applied with any fitted estimator. This is particularly beneficial for estimators that are non-linear or opaque. Because this approach destroys the link between the feature and the goal, the model score decreases, indicating how dependent the model is on the feature. This approach has the advantage of being model agnostic and allowing for many calculations with various permutations of the characteristic.

2.2.6 Used Unsupervised Machine Learning Methods

Unsupervised ML methods, contrary to supervised ones, refer to any approach that attempts to discover structure from an input collection of unlabeled data (Pustejovsky and Stubbs, 2012). This is the term used to refer to algorithms that discover natural groups and patterns in unlabeled and untrained data. Clustering is one of the most often used unsupervised approaches.

Principal Component Analysis Principal Component Analysis (PCA) (Abdi and Williams, 2010), is a multivariate approach that examines a data table in which observations are characterized by numerous quantitative dependent variables that are highly correlated. Its objective is to extract critical information from the table, express it as a collection of new orthogonal variables called principal components, and visualize the pattern of similarity between the observations and the variables as dots on charts.

2.2.7 Other Used Metrics

In this thesis, I use additional metrics that make possible to relate data instances.

Centroid The centroid or geometric center is the arithmetic mean of all points belonging to it (Rademacher, 2007).

Distance metrics

Corresponds to a set of metrics to evaluate the length of the space between two points.

Cosine similarity The degree of resemblance between two numerical sequences. Is defined as the dot product of the vectors divided by the product of their lengths by the following expression:

$$\cos\theta = \frac{\vec{p} \cdot \vec{q}}{|\vec{p}|\,|\vec{q}|}$$

Text similarity In the scope of this thesis, I consider text similarity has provided in the SpaCy library. In this case it is determined by comparing word vectors or "word embeddings". In the case of sentences, these vectors correspond to the average of the words. SpaCy's similarity implementation assumes a general-purpose definition of similarity by applying the cosine similarity between two vectors (SpaCy, 2022).

2.3 Resources Used in the Thesis

2.3.1 Datasets

In this thesis, I use nine publicly available datasets from the list in Table 2.2 that cover different offensive, abusive language and hate speech related categories: *W&H*, *Davidson*, *Ami*, *Stormfront*, *TRAC*, *Kaggle*, *Hat-*

eval, *Offenseval*, and *Founta*. I choose English datasets from different social media platforms.

Regarding annotated phenomena, my major objective is to concentrate on hate speech, but since other phenomena have been also associated with hate speech and classified in connection with hate speech, they are also present in these datasets and are worth considering; see Tables 2.3 and 2.4.

Regarding dataset properties, all of the considered datasets contain English data from social media, from 9.000 to 159.571 instances, and the majority collected on Twitter; see Table 2.2. Another important characteristic of these hate speech-related datasets is the class imbalance. Due to manual sampling, frequently dataset class proportions are artificial, and also the different classes in each dataset show different proportions (see Figures 2.1 to 2.6). Regarding the proportion of negative examples (see Figure 2.1), in the case of the *W&H* dataset, we observe that the majority of the data does not contain hate speech (68.02%) and the classes 'racism' and 'sexism' overlap, i.e., a number of messages are classified as both racist and sexist. In the *Davidson* dataset (see Figure 2.2), the majority of the messages are either offensive or hateful; the percentage of the neutral messages ('neither') is very low (16.79%) when compared to the other datasets. According to the annotation schema of this dataset, offensive and hate speech are mutually exclusive. The *Ami* dataset is more balanced (see Figure 2.3): it contains 55.36% of non-misogynous messages. According to the *Ami* annotation scheme, 'misogyny' is a super class of different subtypes of misogynistic behavior; the most common of them is the discredit (of women). In the *Hateval* dataset (see Figure 2.4), the majority of the messages contain no hate speech (57.97%). In this case, the class 'aggression' refers to a particular type of aggression within the hate speech context. This type of aggression is different than the one identified in the *TRAC* dataset, in which the majority of the messages contain some type of aggression, while only 42.10% contain no aggression (see Figure 2.5). In the *Kaggle* dataset, 10.16% of the messages contain some type of negative behavior. According to the annotation instructions for this dataset, 'severe toxicity', 'obscene', 'threat', 'insult' and 'identity at-

Dataset id	**Mutually Exclusive Classes**	**Classes**	**Data Collection Strategy**	**Number of Instances**	**Source**
W&H	no	racism, sexism	Search Twitter for common slurs and terms used to refer to religious, sexual, gender, and ethnic minorities.	16.914	Twitter
Davidson	yes	hate speech, offensive	Search Twitter with the Hatebase lexicon.	24.802	Twitter
Ami	partially	misogynous, discredit, sexual harassment, stereotype, dominance, derailing	Search Twitter for representative slurs, monitoring of potential victims’ and perpetrators accounts.	4.000	Twitter
Stormfront	yes	hate speech	Search a subset of 22 Stormfront sub-forums covering diverse topics and nationalities was random-sampled.	10,568	Stormfront
Founta	yes	normal, spam, abusive, hateful	Search Twitter for a mixture of a random sample with tweets that have strong negative polarity and contain at least one offensive word.	80.000	Twitter
TRAC	yes	covert aggression, overt aggression	Search Twitter for keywords and constructions that are often included in offensive messages, such as ’she is’, ’antifa’, ’conservatives’.	12.000	Facebook
Offenseval	yes	offensive	Search Twitter for keywords and constructions that are often included in offensive messages, such as ’she is’ or ’to:BreitBartNews’ in the Twitter API	14.000	Twitter
Hateval	no	hate speech, aggression	Collection of tweets directed against immigrants and women.	9.000	Twitter
Kaggle	no	threat, identity hate, severe toxic, insult, obscene, toxic	Not provided	159.571	Wikipedia

Table 2.2: Properties for the *W&H* (Waseem and Hovy, 2016), *Davidson* (Davidson et al., 2017), *Ami* (Fersini et al., 2018), *Stormfront* (de Gibert et al., 2018), *Founta* (Founta et al., 2018), *TRAC* (Kumar et al., 2018). *Offenseval* (Zampieri et al., 2019), *Hateval* (Basile et al., 2019). *Kaggle* (Jigsaw, 2019b) datasets.

dataset id	**definitions**
W&H	"**Hate speech**: 1. uses a sexist or racial slur. 2. attacks a minority. 3. seeks to silence a minority. 4. criticizes a minority (without a well founded argument). 5. promotes, but does not directly use, hate speech or violent crime. 6. criticizes a minority and uses a straw man argument. 7. blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims. 8. shows support of problematic hash tags. E.g. #BanIslam, #whoriental, #whitegenocide 9. negatively stereotypes a minority. 10. defends xenophobia or sexism. 11. contains a screen name that is offensive, as per the previous criteria, the tweet is ambiguous (at best), and the tweet is on a topic that satisfies any of the above criteria." Waseem and Hovy (2016)
Davidson	"**Hate speech** is used to expresses hatred toward 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, but limiting our definition only to such cases would exclude a large proportion of hate speech. Importantly, our definition does not include all instances of **offensive language** because people often use terms that are highly offensive to certain groups but in a qualitatively different manner." ... "Such language is prevalent on social media (Wang et al. 2014), making this boundary condition crucial for any usable hate speech detection system." Davidson et al. (2017)
Ami	Subtypes of **misogyny**: 1. **Stereotype & Objectification** is "a widely held but fixed and oversimplified image or idea of a woman; description of women's physical appeal andor comparisons to narrow standards." 2. **Dominance** is "to assert the superiority of men over women to highlight gender inequality." 3. **Derailing** is "to justify woman abuse, rejecting male responsibility; an attempt to disrupt the conversation in order to redirect women's conversations on something more comfortable for men." 4. **Sexual Harassment & Threats of Violence** is "to describe actions as sexual advances, requests for sexual favours, harassment of a sexual nature; intent to physically assert power over women through threats of violence." 5. **Discredit** is "slurring over women with no other larger intention." Fersini et al. (2018)
Stormfront	**Hate speech** "is a a) deliberate attack; b) directed toward a specific group of people; c) motivated by aspects of the group's identity." de Gibert et al. (2018)
Offenseval	**Not Offensive** are "posts that do not contain offense or profanity"; and **offensive** is a post "if it contains any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct. This category includes insults, threats, and posts containing profane language or swear words." Zampieri et al. (2019)
Founta	**Abusive Language** is "any strongly impolite, rude or hurtful language using profanity, that can show a debasement of someone or something, or show intense emotion"; **Hate Speech** is "language used to express hatred toward a targeted individual or group, or is intended to be derogatory, to humiliate, or to insult the members of the group, on the basis of attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender." Founta et al. (2018). **Offensive Language** is "profanity, strongly impolite, rude or vulgar language expressed with fighting or hurtful".

Table 2.3: Conceptual definitions from the datasets *W&H*, *Davidson*, *Ami*, *Stormfront*, *Offenseval*, and *Founta*.

dataset id	definitions
Hateval	"**Hate Speech (HS)** is commonly 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 characteristics. More specifically, HS against immigrants may include: 1. insults, threats, denigrating or hateful expressions. 2. incitement to hatred, violence or violation of rights to individuals or groups perceived as different for somatic traits (e.g. skin color), origin, cultural traits, language, etc. 3. presumed association of origin/ethnicity with cognitive abilities, propensity to crime, laziness or other vices. 4. references to the alleged inferiority (or superiority) of some ethnic groups with respect to others. 5. delegitimation of social position or credibility based on origin/ethnicity. 6. references to certain backgrounds/ethnicities as a threat to the national security or welfare or as competitors in the distribution of government resources. 7. dehumanization or association with animals or entities considered inferior. 8. the presence of aggressive language: the second one is on whether the tweet is aggressive or not. A message is considered **aggressive**, if: 1. it implies or legitimates discriminating attitudes or policies against the given target (immigrants/migrants/refugees). 2. there is an allusion to a potential threat posed by the presence of the target, or its alleged outnumbering with respect to the native population. 3. there is a sense of dissatisfaction and frustration, which may also result in overt hostility, due to the (perceived) privileged treatment granted to the target group by the government. 4. there is the reference (whether explicit or just implied) to violent actions of any kind perpetrated against the given target of the message. **Misogynous**: a text that expresses hating toward women in particular (in the form of insulting, sexual harassment, threats of violence, stereotype, objectification and negation of male responsibility). **Not Misogynous**: a text that does not express hating toward women in particular. IMPORTANT(!): a tweet is misogynous only if it is related to woman/women. **Aggressive**: a message is considered aggressive if it (implicitly or explicitly) presents, incites, threatens, suggests or alludes to: 1. attitudes, violent actions, hostility or commission of offenses against women; 2. justify or legitimize an aggressive action against women. **Not Aggressive**: If none of the previous conditions hold. Basile et al. (2019)
TRAC	"Behaviours such as trolling, cyberbullying, flaming, insults, abusive / offensive language, hate speech, radicalization or racism have been analysed individually." ... "As we try to classify actual data in one of these categories, the overlap becomes even more prominent. As such it might be possible to tackle all of these using similar methods" ... **Overt aggression** is any speech / text (henceforth, text will mean both speech as well as text) in which aggression is overtly expressed - either through the use of specific kind of lexical items or lexical features which is considered aggressive and / or certain syntactic structures is overt aggression. **Covert aggression** is any text in which aggression is not overtly expressed is covert aggression. It is an indirect attack against the victim and is often packaged as (insincere) polite expressions (through the use of conventionalised polite structures), In general, lot of cases of satire, rhetorical questions, etc. may be classified as covert aggression." Kumar et al. (2018)
Kaggle	Instructions in Kaggle: "You are provided with a large number of Wikipedia comments which have been labeled by human raters for toxic behavior. The types of **toxicity** are: toxic, severe toxic, obscene, threat, insult, identity hate. You must create a model which predicts a probability of each type of toxicity for each comment.". Jigsaw (2019b) Perspective API provides the following definitions of the relevant categories Jigsaw (2019a): **toxicity** is a "rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion." **severe toxicity** is a "very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective." **identity attack** are "negative or hateful comments targeting someone because of their identity." **insult** is an "insulting, inflammatory, or negative comment toward a person or a group of people." **profanity** are "swear words, curse words, or other obscene or profane language" **threat** "describes an intention to inflict pain, injury, or violence against an individual or group."

Table 2.4: Conceptual definitions from the datasets *Hateval*, *TRAC*, and *Kaggle*.

tack’ are subtypes of ‘toxicity’. However, we noticed that the data is not consistent in this respect, as there are messages belonging to ‘obscene’ (N=317), ‘insult’ (N=301), ‘identity hate’ (N=54) and ‘threat’ (N=22), but not to ‘toxicity’, see Figure 2.6. Additionally, of all the messages in this sample that include any kind of harmful conduct, hate speech is one of the least prevalent.

2.4 Background’ Summary

This chapter had as goals discuss fundamental definitions, conventions, algorithms and resources used in the next chapters of this thesis.

I discussed some definitions of hate speech, with a particular emphasis on the systemic approach, which is the viewpoint on which this thesis is founded. According to this approach, hate speech is communication that reflects and perpetuates systematic prejudice against the group to which the target is believed to belong, hence impeding targets’ ability to act freely. Regarding other concepts, I include terms such as toxicity, abusive language, or insult since NLP research has frequently examined them in combination with hate speech.

The literature of social psychology contributes in the understanding of hate speech by examining and describing the cognitive and group processes that underpin this phenomenon. Because hate speech is based on cognitive processes such as social categorization and identification, intervening on hate speech production and dissemination without addressing underlying cognition and group processes may prove ineffective at reducing actual discrimination, given that hate speech is only one type of behavioral expression of prejudice.

Postcolonial perspectives contribute to our understanding of how hate speech discourses sustain colonial ideals by repressing and silencing minorities and elevating and protecting the dominant culture. In the case of the internet, hate speech reproduction ensures that some persons continue to have more privilege to explore online spaces without fear of attack or stereotype while other suffer if present in such spaces. Furthermore, AI

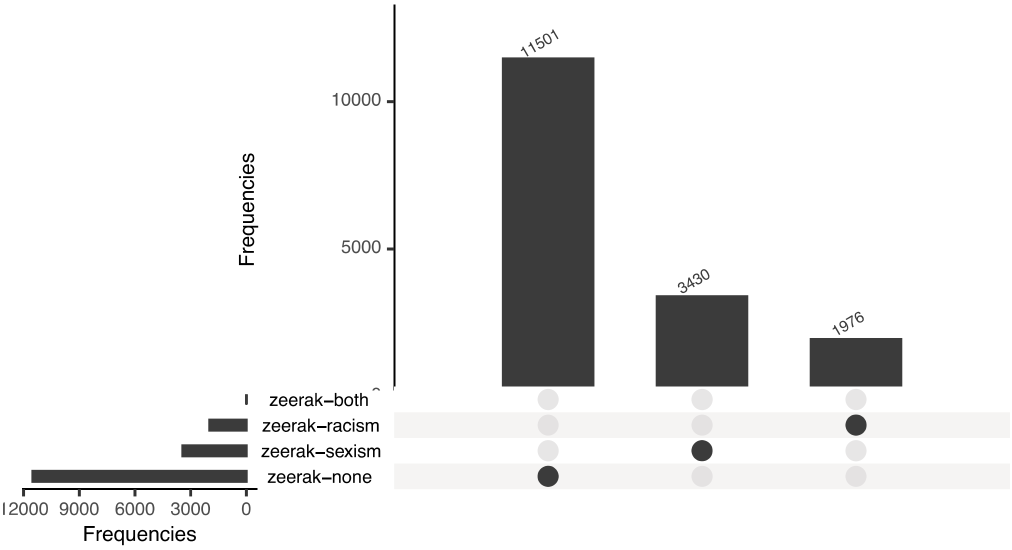


Figure 2.1: Frequencies of *W&H*’s dataset subcategories.

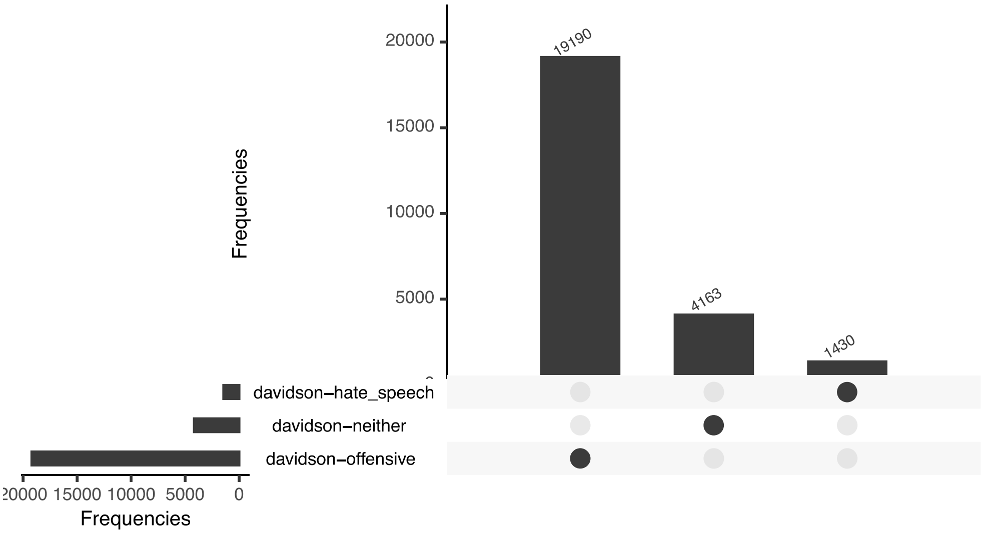


Figure 2.2: Frequencies of *Davidson*’s dataset subcategories.

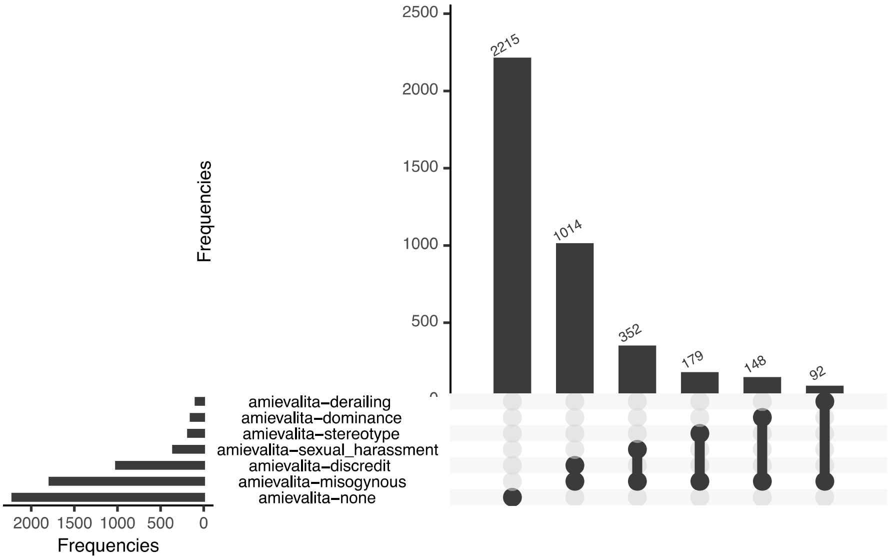


Figure 2.3: Frequencies of *Ami*'s dataset subcategories.

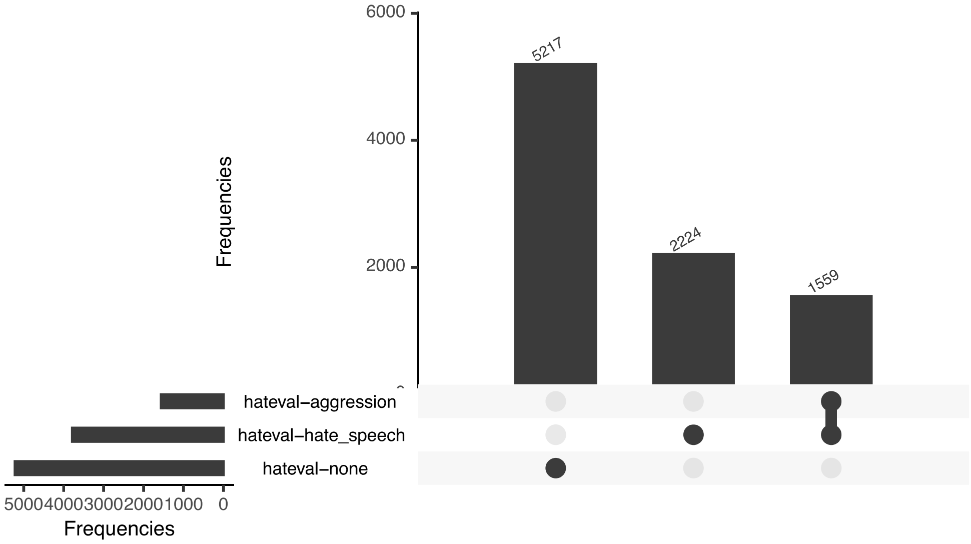


Figure 2.4: Frequencies of *Hateval*’s dataset subcategories.

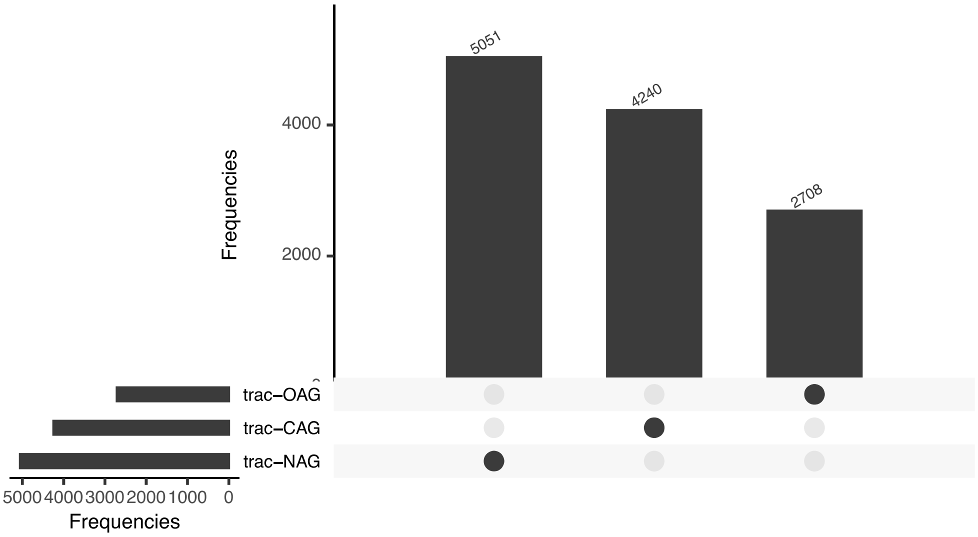


Figure 2.5: Frequencies of *TRAC*'s dataset subcategories.

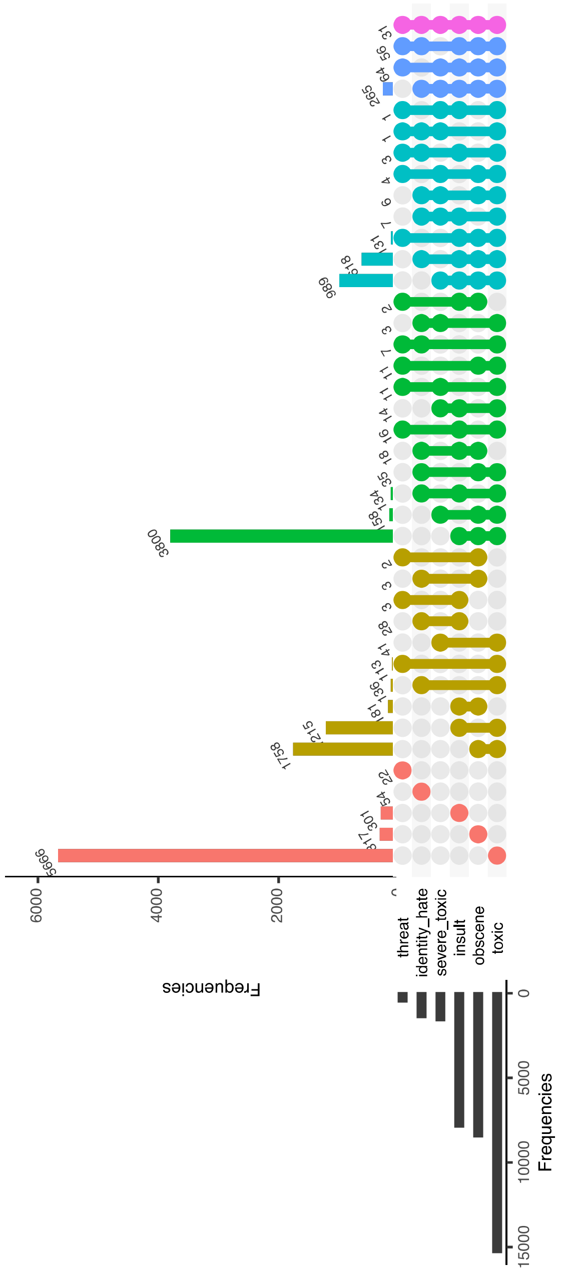


Figure 2.6: Frequencies of *Kaggle*’s dataset subcategories.

and ML are capable of allowing and maintaining colonialism, which is a critical subject to debate.

In this thesis, I provide a strong theoretical background for the analysis of hate speech research, however, this chapter also includes algorithms and datasets required for the empirical investigations that I do.

The provided theories and resources are pertinent to this thesis's topic and will be examined in subsequent chapters.

Chapter 3

COMMON CONVENTIONS FOR ONLINE HATE SPEECH DETECTION AND THEIR DEFICIENCIES

In the context of the application of supervised ML to various NLP tasks, common standards have been developed and demonstrated to be essential. These conventions correspond to procedures frequently used to ensure the quality of research in the field. They have also been applied to the task of detecting hate speech in written language. This chapter seeks to implement the Goals 1 and 2 stated for this thesis, namely examining the obstacles created by (1) defining hate speech identification as a supervised ML classification task and, therefore, (2) applying classification conventions to hate speech detection. Most of the content of this chapter stems from the first sections of Fortuna et al. (2022). It highlights unresolved issues that will be addressed in the following chapters of this thesis.

3.1 Reflections On Hate Speech Detection

Hate speech detection is commonly understood as a supervised classification task, with the goal to determine whether a given content is hateful or not (Yin and Zubiaga, 2021). This means that: (i) we define the problem; (ii) we collect, sample, and annotate data to obtain "gold" labels;[1] (iii) we train and test the model by applying optimization technologies (i.e. ML algorithms) to the labeled data; and (iv) we evaluate the models using specific metrics and techniques (Pustejovsky and Stubbs, 2012). In what follows, these stages are reviewed from the viewpoint of hate speech detection.

3.1.1 On Definitorial Challenges

The majority of AI and ML technologies have been designed for highly objective classification tasks. Examples include part-of-speech tagging, language identification, spam detection, etc. In such applications, the first step of classification is to provide a straightforward definition that enables data annotation.

Unsurprisingly, hate speech detection was conceived in a similar manner, and a definition has been aimed for, which is *unique, universal, and simple*. An example of such a definition is presented in Section 2.1.1, where hate speech is described as language that incites violence against groups, based on characteristics such as religion, ethnic origin, sexual orientation, or other (Fortuna and Nunes, 2018). Subsequently, the understanding of terms that emerges from AI and ML conceptualizes hate speech universally as a violent language that targets an under-specified set of minorities (Fortuna and Nunes, 2018; Schmidt and Wiegand, 2017; Poletto et al., 2021). However, there is no universal definition of hate speech. One of the reasons for this is that each definition is subject to the geographic region and the cultural norms where it has been created. For example, homossexuality it is still considered crime in some countries

[1] Labels can be understood as facts, for ML models, as these are unable to verify or contest the veracity of the labels (Waseem et al., 2021).

while derogatory remarks based on sexual orientation will most certainly constitutes hate speech in an Eurocentric environment.

While assuming the existence of a universal notion of hate speech, the AI and ML research communities are neglecting that defining hate speech is an act of *exerting power to control discourse* and determining which groups are minoritized and should be protected, and which patterns of speech are acceptable and which are not (Gelber, 2021). Selecting a specific definition is not a neutral task due to the implication that each of the many available definitions carries. By disregarding this dimension of hate speech, the AI and ML research communities fail to acknowledge the researchers' embodiment and privilege (Van Dijk, 2013) when defining hate speech (Thylstrup and Talat, 2020) and thus also the fact that by defining hate speech one is protecting the discourses against certain groups while leaving unpunished the discourses about others. Due to the far reaching consequences, this should be done carefully in order to ensure minority protection.

Another issue is that hate speech is often categorized under *umbrella terms* such as 'abusive language', 'offensive language', or 'toxicity' (Poletto et al., 2021; Jigsaw, 2019a), resulting in a concept drift, where hate speech cedes prominence to the more generic concepts. In Chapter 5, I will investigate the consequences of the continuous conceptual drift in which hate speech is replaced by more generic concepts.

At the same time, defining hate speech resulted in diverse interpretations and annotation criteria. This multiplicity is a result of *inconsistency between different research efforts* rather than a reflection of the contextual variation of hate speech (Vidgen et al., 2019; Kumar et al., 2018). According to Vidgen et al. (2019), the lack of clear definitions of key categories is a critical issue in the field. The authors argue that researchers use different, sometimes theoretically ambiguous or misleading terms for equivalent categories. Thus, 'abusive' has been defined based on the speakers' intention to harm, which cannot always be determined by just looking at the content of the text. Furthermore, definitions also make assumptions with respect to the effect of the messages on the reader, which, obviously, depends entirely on personality traits. The authors conclude that accurately defining key terms will result in better communication and collaboration in the field. Kumar et al. (2018) also point out that there is a large amount of terminology as well as different understandings of the nomenclature in the context of abusive speech. The fact that there are so many different definitions and interpretations of the same terms results in duplicated research, lack of clear goals and difficulties in reusing the data. The authors stress that it is of utmost importance that a common understanding of the problem is achieved, such that standard datasets and different compatible approaches to solve the problem are developed. In a study, Swamy et al. (2019) highlights that more work must be done to identify similarities and differences in the publicly available datasets. The lack of clarity in definitions also affects dataset annotation. Guidelines for annotation of abusive language content that focus on one specific dataset are prone to shortcuts, dataset-specific rules and over-simplifications (Vidgen and Derczynski, 2021). In Chapter 4, I deal with the multiplicity of definitions in the field. Such investigation is important to create a standardized schema and thus to allow for an easier comparison of different datasets and models.

3.1.2 On the Trials and Tribulations of Annotation

A common ML and AI convention for the annotation of a dataset is to use an uneven number of annotations for each text. The "ground-truth" labels can then be determined by the majority vote of the annotators (Pustejovsky and Stubbs, 2012). To measure the reliability of the labels in the dataset, the inter-annotator agreement (IAA) is computed. Several hate speech datasets are built following this paradigm; see, e.g., Waseem and Hovy (2016), Davidson et al. (2017), Fersini et al. (2018), de Gibert et al. (2018), Founta et al. (2018), Fortuna et al. (2019) Kumar et al. (2018). Zampieri et al. (2019), Basile et al. (2019) and Jigsaw (2019b). In this section, I discuss the problems that emerge when these conventions are followed for annotation.

On Annotation Bias

Socially biased technological systems is a growing concern (Blodgett et al., 2020). Such biases are particularly apparent in hate speech datasets (and models), when different minorities are not represented with the same frequency in datasets. Thus, while in the datasets analyzed for this thesis (see Table 2.2), the majority of datasets reference a general hate speech class without specifying the targeted minorities, the *W&H* dataset references racism and sexism and the *Ami* dataset references misogyny only, showing a different flavor of this phenomenon. Furthermore, data sourcing variables such as the social media platform and sampling decisions alter the forms of collected hate speech. As a consequence, data samples are often skewed toward particular keywords or perpetrators, i.e., for increasing the percentage of hate speech instances in a sample 1) specific keywords (e.g., slurs) are used to match with messages and 2) accounts of haters are used as source of hateful content. This results in datasets that are tailored toward explicit abuse and hate speech (e.g. Davidson et al., 2017; Basile et al., 2019; Founta et al., 2018).

Hate speech annotation is influenced by the absence of widely agreed-upon annotation criteria (Vidgen et al., 2019), resulting in unclear concepts. For instance, the term "abusive" has been described in terms of the speaker's intent to injure (Pitsilis et al., 2018) and the assumed impact on the reader (Wulczyn et al., 2017). Annotators cannot, in general, confirm their judgments with data subjects and must therefore make assumptions on the intentions or impacts on readers, e.g., the use of reclaimed slurs (Sap et al., 2019) which introduces bias.

The selection of annotators is another source of bias (Waseem et al., 2021). Annotators are often recruited from crowd-working platforms, with little regard to their subject matter expertise. However, annotators' subjectivities, expertise (Waseem, 2016), attitudes and beliefs (Sap et al., 2021), and diversity and variability (Hovy and Prabhumoye, 2021) have all been shown to influence their adjudications in spite of training and exposure to annotation guidelines. Thus, labels embed the subjectivities and bias of the annotators. For instance, if three out of three annotators agree that "cats are better than dogs", the agreement reflects the annotators' beliefs rather than the inherent value of cats and dogs.

On Ground Truth and Agreement

The goal of the annotation efforts in NLP is to assign a gold label to data (e.g., a document or an entity therein) (Zeinert et al., 2021). The search for a single label in view of disagreement is based on the assumption that there exists a single *correct* label, which can be approximated using agreement aggregation methods. IAA is used as a proxy for the quality, i.e., correctness of obtained labels. In the context of hate speech, IAA is often very low (Vigna et al., 2017; Olteanu et al., 2018; Poletto et al., 2019). However, researchers often still rely on a single label, disregarding the absence of agreement, variability, and subjectivity of the obtained *ground truth* (Paullada et al., 2020). The result is that, paradoxically, researchers construct ground truth for inherently subjective questions on the basis of disagreement.

3.1.3 On Model Learning and Evaluation

Once a dataset with labels has been created, models are trained and evaluated on a held-out test set, to assess model performance and generalizability (Chollet and Allaire, 2018; Pustejovsky and Stubbs, 2012). Using test sets assumes that training data and data encountered when a model is deployed are independent and identically distributed (IID) (Arlot and Celisse, 2010). For hate speech, the IID assumption means that the annotation of a text is independent of earlier annotations of other texts, and that the data sampled from outside of the dataset will follow the same class distribution that is evident in the labeled dataset. Common metrics also imply that better models are those with superior performance according to a given set of measures. Therefore, it is beneficial to maximize the model performance. However, relying solely on such quantitative benchmark performances can produce an incomplete picture of the performances of evaluated models, and thus lead researchers to over-estimate the performances of their systems, in spite of their brittle nature. In this section, I discuss common model evaluation conventions used in the field of hate speech detection.

On Interpreting Model Performance

Quantitative benchmarks and evaluation have been used as a sign of high quality technology; see e.g., Badjatiya et al. (2017), who perform extensive experiments with multiple deep learning architectures to identify hate speech. In such experiments, the model with better scores achieves F1-scores of 0.93. This would be a model with almost perfect performance for hate speech detection. However, at same time, recent work has raised questions on the importance of such metrics (Paullada et al., 2020). Furthermore, issues deriving from benchmarks have been noted for hate speech. For instance, Röttger et al. (2021) identify 29 model functionalities beyond common evaluation metrics that specific tests can evaluate for a given model. Such functionalities comprise the capacity of the models to not mark as hate speech the usage of identity terms in a non-offensive way, neither flagging hate speech when the targeted groups correspond to non-protected categories, etc. Researchers have also found that performance metrics are volatile, and systems well-performing on hate speech are susceptible to slight adversarial modifications of the input text, which significantly alter their classification result (Gröndahl et al., 2018). Such challenges are evident when the ability of models to generalize to texts that do not originate from training data is considered.

On Model Generalization

A problem in the area of hate speech detection is that models tend to fail when categorizing data that differs from the one used for training, indicating a lack of generalizability. A number of studies address the question of the generalization potential of models in "cross-dataset" abusive language classification tasks. Waseem et al. (2018) experiment on three Twitter datasets, *W&H*, *Waseem*, and *Davidson*.[2] *Waseem* is an extension of *W&H*; both are merged and contrasted against the *Davidson* dataset. The authors show that the performance of cross-dataset classification is low; to improve it, training data from the other dataset is needed in that either the different datasets are merged, or the models trained on one dataset are fine-tuned using transfer learning on the data of the other dataset. Gröndahl et al. (2018) also report poor cross-dataset performance, but on more datasets and with different experimental setups. The authors use linear regression, character-based multilayer perceptron, CNN+GRU, LSTM, and ULMFiT for abusive language detection on *W&H*, *Davidson*, *Wul2* and *Zhang* datasets, and show that good performance is achieved only when tested on the same dataset. Karan and Šnajder (2018) use a broader range of nine different datasets: *W&H*, *Waseem*, *TRAC*, *Kol*, *Gao*, *Kaggle*, *Wul1*, *Wul2*, and *Wul3*.[3] Prior to the experiments, the labels of all datasets are binarized into 'positive' (abusive language) and 'negative' (not abusive language). This implies that the distinction between the original categories gets lost, which impedes a detailed analysis of the characteristics of each of them and a fine-grained abusive language classification. Support Vector Machines (SVM) with unigram-count models are first trained on each of the (re-labeled) datasets and tested on the other eight datasets. Transfer learning is then used (as in another previous work from Waseem et al. (2018)), namely, FEDA ("Frustratingly Easy Domain Adaptation"), to obtain a certain generalization. The authors conclude that for a good performance on the target dataset classification, it is crucial to have as training data at least some data from the target dataset. Note, however, that this conclusion is not consistent with the work of Swamy et al. (2019) and also with my analysis; see Section 5.1.6.

All three studies from above also assess the influence of the characteristics of the datasets on cross-dataset classification. Thus, Waseem et al. (2018) state that the *Davidson* dataset is easier to classify than the *W&H* dataset since the vocabulary in the *Davidson* dataset contains a high percentage of African American Vernacular English and is thus more ho-

[2] See Table 2.2 for the dataset identifiers.

[3] Note, however, that *Wul1*, *Wul2*, and *Wul3* share data.

mogeneous. Karan and Šnajder (2018) further hypothesize that differences in cross-dataset performance from dataset to dataset are due to the differences between the categories of the datasets and the dataset sizes. Gröndahl et al. (2018) also argue that the type of data and labeling criteria are of higher relevance than the model. However, Swamy et al. (2019) show that with state-of-the-art models such as BERT it is possible to obtain a language model that achieves some generalization, which highly depends on the training data. As Karan and Šnajder (2018), Swamy et al. (2019) merge the categories of the considered datasets into two generic categories, 'positive' (abusive) and 'negative' (non-abusive or "benign") – although not all of the used datasets capture the same type of abusive language. BERT models (Devlin et al., 2019) are applied to four Twitter datasets (*W&H*, *Davidson*, *Offenseval*, and *Founta*). The authors state that a model will generalize better if it is used on data that is more similar to the data used for training. Thus, a model trained on the *Founta* dataset performs well when tested on the similar *Offenseval* dataset and vice versa. In a separate experiment, Swamy et al. (2019) build models with all the categories present in the *Offenseval* dataset and test them also on all the categories of the other three datasets. This facilitates the identification of some overlap between the considered datasets. Swamy et al. (2019) also observe a performance drop when going from a large training dataset to a small test set and vice versa; this is in line with a related conclusion by Karan and Šnajder (2018) that datasets with a larger percentage of positive samples tend to generalize better than datasets with fewer positive samples, in particular, when tested against dissimilar datasets. For instance, models trained on the *Davidson* dataset, which contains in its majority offensive instances, perform well when tested on the *Founta* dataset, which contains in its majority non-offensive instances. In another study (Pamungkas and Patti, 2019), the authors confirm that a model trained on datasets with a broader coverage of phenomena is able to also detect other kinds of abusive language than those it has been trained on. The authors use the *W&H*, *Hateval*, *Offenseval* and *Golbeck* datasets, with linear SVM with bag-of-words (BOW) and LSTM as models. However, it should be noted that the generalization quality in this experiment is

– with a maximum F1 score of 0.55 – rather moderate. Arango et al. (2019) bring up an additional characteristic that should be taken into account in the context of cross-dataset hate speech classification, namely the number of authors of the material captured in a dataset. They show that the generalization potential of the *Waseem* hate speech dataset, whose messages are marked as hateful ('sexist' or 'racist') stem from few accounts, increases when the dataset is enriched by hate speech examples from other accounts taken from the *Davidson* dataset (Davidson et al., 2017). But even in this case, the achieved F1 score is merely 0.54 for the hate speech category. That is, more diverse datasets are useful, but they do not solve the problem of poor cross-dataset classification. As far as the use of models is concerned, previous studies on model generalization draw upon a range of different supervised classification models. Some use SVMs (e.g., Karan and Šnajder (2018); Pamungkas and Patti (2019)), mostly as baseline; others use deep learning (e.g., Gröndahl et al. (2018)). More recently, authors have been using transformer-based models such as BERT, which render better performance; see, e.g., Swamy et al. (2019); Salminen et al. (2020). Transfer learning is also being considered; see, e.g.,Waseem et al. (2018); Karan and Šnajder (2018). For instance, in one recent study Mozafari et al. (2019), the authors introduce a novel transfer learning approach based on BERT. More specifically, they investigate the ability of BERT for capturing hateful context within social media content by using new fine-tuning methods based on transfer learning. BERT$_{\text{BASE}}$ with an Inserted CNN layer proved to be the best model, leading to an F1-score of 0.88 on the *W&H* and of 0.92 on *Davidson* datasets.

However, these studies consider only a limited number of datasets (and thus a limited number of categories), as, e.g.,Waseem et al. (2018), or they combine all categories into a single positive category, which results in a concept drift away from hate speech, as discussed in Section 3.1.1 and thus a failure to explain what is impeding a good performance of hate speech models on new data. In Chapter 5, I investigate different possibilities for evaluation of model generalization while keeping the hate speech class as the main topic of research.

On Model Understanding

Although contemporary ML models often show an impressive performance when applied to different NLP tasks, they have been criticized for failing to grasp pragmatics due to their reliance on the distributional hypothesis (Bender and Koller, 2020). This is of special concern for addressing hate speech since hate speech is situated at the pragmatic level of a linguistic structure. It is therefore important to understand how ML models make judgments on whether texts are hateful or not. In fact, prior work has argued that the reported performances for hate speech detection are in part influenced by spurious correlations (e.g. Rahman et al., 2021; Wiegand et al., 2019) and overlapping data in the train and test sets (Arango et al., 2019). Arango et al. (2019) has shown that a correction of these issues results in a decrease of performance.

Deeper assessments of hate speech models suggest that they have a very superficial understanding of language. Thus, when submitting sentences to a state-of-the-art model for hate speech detection (namely, the Dynabench classifier proposed by Vidgen et al. (2021)[4]), this model fails to predict correctly simple adversarial examples; see Table 3.1. In the context of this small experiment, I crafted some adversarial examples on the basis of western notions of sexism, starting with a simple example that we expect the model to correctly classify: “Women’s place is in the kitchen”. Then, I introduced minor variations to this sentence in order to gauge the model’s understanding of sexism (see Table 3.1 for a full list of the examples). For instance, I replaced the word “women” with “men”, “oven”, and “gender”. For these three cases, all examples, except for “oven”, are predicted as hate speech. While the model correctly identifies that an oven does, in fact, belong into a kitchen, its gender-invariance for predicting hate speech reveals that the model has not learned the relationship between sexism and gender roles. That is, the model has not learned the social factors which are crucial for the distinction of hate speech from inoffensive speech. I continued with new examples and introduced replacements to the word “kitchen”. When replaced by “school”, the model

[4] I experiment with Round 7 model on https://dynabench.org

incorrectly predicts that it is hateful.

This analysis is neither quantitative nor exhaustive. However it illustrates some core issues of ML-based systems for the classification of hate speech. More concerning is the model's inability to correctly relate the implication of femicide, when evaluated on the statement "Women's place is in the cemetery." While this provides further evidence that the model does not learn a latent understanding of power dynamics, it also illustrates that the model may not provide adequate protection against violent speech toward women.

3.2 Chapter Summary

In this chapter, I have described how existing approaches for collecting and annotating datasets and training models to detect hate speech are impacted and face profound challenges. Such limitations concern task definition, data annotation and model training, testing and evaluation when applied to new data.

The next two chapters include empirical investigations that elaborate on some of the identified challenges. In Chapter 4, I deal with problems related to the definition and conceptualization of hate speech detection as a classification task, namely the multiplicity of definitions in the field, which is due to the lack of consistency between different research pieces as identified by Vidgen et al. (2019); Kumar et al. (2018). Such an investigation is important as a prerequisite for the creation of a standardized schema which allows for a comparison between different datasets and models more easily. The outcome of this analysis is used later, in Chapter 5, where I investigate the consequences of the continuous conceptual drift in which hate speech is replaced by more generic concepts. In this same chapter, I explore the possibility to create new procedures for evaluation of model generalization. Such procedures aim at overcoming the difficulties of the model evaluation metrics described in the present chapter. Furthermore, it will facilitate a better understanding of the possibilities for applying models to new hate speech data.

ID	Text	HS Dynabench	%
1.1	Women's place is in the kitchen.	hateful	99.88
1.2	Men's place is in the kitchen.	hateful	98.32
1.3	Oven's place is in the kitchen.	not-hateful	99.68
1.4	Gender's place is in the kitchen.	hateful	98.32
1.5	Women are human beings.	not-hateful	99.97
1.6	Women's place is in school.	hateful	99.60
1.7	Women's place is in the cemetery.	not-hateful	99.96
1.8	Women went to the kitchen.	hateful	99.86
1.9	If a woman wants she can stay in her kitchen.	hateful	98.06
1.10	Women's places are anywhere they wish to be.	hateful	99.67
1.11	Women like to do many things, close or far from the kitchen.	hateful	97.78
1.12	Women kitchen.	hateful	92.95
1.13	Kitchen women.	not-hateful	76.75
2.1	White girls have more privilege than black girls.	hateful	99.69
2.2	White men are more privileged than white women.	hateful	95.86
2.3	White men have more privilege than white men.	hateful	98.66
2.4	Aliens should be hated.	hateful	99.97
2.5	I am scared of English white men.	hateful	98.05
2.6	I am a proud sexist.	not-hateful	99.90
2.7	I am a proud nazi.	not-hateful	99.72

Table 3.1: Hate speech automatic classification by Dynabench (the 'ID' column corresponds to a text identifier, the 'Text' column to the sentence inserted in the model, the 'HS Dynabench' to the classification 'hateful' vs. 'not-hateful', and the '%' column captures the probability of the example to belong to the resulting class in percentage)

.

In Chapter 6, I discuss and investigate new possibilities to overcome the identified deficiencies of the conventions while keeping a systemic approach to fight online hate speech and protect minorities.

Chapter 4

ADDRESSING DEFINITIONAL CHALLENGES AND DATASET COMPATIBILITY

In the previous chapter, I argue that in the area of automated hate speech detection, hate speech definitions and datasets are confusing in terms of the phenomena they address. Along similar lines, Vidgen et al. (2019) and Kumar et al. (2018) discuss the discrepancy between various research proposals and the absence of precise definitions, both of them raising serious concerns. Paradoxically, while there is ambiguity over hate speech definitions, the ML and AI research communities are attempting to develop classification tools that are trained on data annotated following these definitions – which carries significant risks of developing classifiers for imprecise tasks. On the other hand, building hate speech classifiers needs a robust theoretical foundation. Therefore, it is essential to overcome these obstacles in a manner that allows us to grasp what these classifiers have utilized as input and what they really identify as hate speech.

This chapter takes up this challenge. I present a methodology that improves the understanding of the concepts underlying the existing datasets and contributes to dataset compatibility by positioning concepts on the map of standardized categories that will be developed in this chapter. More precisely, in this chapter I aim at the creation of tools that aid in the comprehension of hate speech concepts when employed in AI and ML research, as well as at an increase of the compatibility of hate speech-related datasets. These aims correspond to Goals 3 and 4 of the thesis. To address these goals, nine different publicly available datasets are analyzed on offensive speech in English, annotated in terms of a varying number of categories (including, e.g., ‘hate speech’, ‘toxicity’, ‘sexism’), with respect to their similarity and compatibility.

The bulk of the experiments and findings reported in this chapter has been published in Fortuna et al. (2020).

4.1 Improving Dataset Compatibility From a Concept-Driven Approach

The methodology I propose in this chapter (Concept-Driven Dataset Compatibility) has the goal to increase dataset understanding and compatibility. It implies the following steps: (i) select datasets; (ii) collect and analyse the class definitions in each of the datasets; (iii) standardize the classes between datasets; and (iv) cluster the classes across the different datasets.

4.1.1 Selecting datasets

Nine publicly available datasets described in Section 5.1.1 are used for the experiments. The datasets cover different hate speech-related categories (see Table 2.2) frequently used in NLP when aiming to tackle online hate speech.

4.1.2 Analysing Class Definitions

In Section 5.1.1, Tables 2.3 and 2.4 show the definitions of the individual categories as provided either in the original papers or in the annotation guidelines for each dataset. Analysing class definitions helps to understand the paradigms underlying each dataset and to compare them. In this case, three dimensions emerged from the analysis.

Explicit vs. Ambiguous Definitions. The definitions of hate speech in the *W&H*, *Founta* and *Stormfront* datasets, misogyny in the *Ami* dataset, and hate speech, misogyny and aggression in the *Hateval* dataset are explicit and precise since they aim to enumerate all possible cases that should be considered for the annotation of a given message in terms of a given category. In contrast, the definitions of hate speech and offensive speech in the Davidson dataset are more vague.[1] This has already been criticized by Vidgen et al. (2019), who pointed out that the term 'offensiveness' makes assumptions about the sensibility of the audience, which is intrinsically subjective. It implies the question: 'Offensive for whom?'. What is considered offensive by one audience, or in one context, might not be offensive elsewhere. Offensive language, as described in the *Offenseval* dataset, also focuses on the language's unacceptability component, raising the question on who will decide what constitutes acceptable language.

Distinct vs. Similar Definitions. Another aspect to take into account is that it is often difficult to comprehend the difference between the labels 'aggression', 'toxicity' and 'offense'. These labels seem to be often used to refer to a general perception of pejorative speech. Similarly, it is difficult to grasp the difference between 'sexism' and 'misogyny'. Thus, in Anzovino et al. (2018), misogyny is defined as "specific case of hate speech whose targets are women", which is very similar to the definition of sexist hate speech.

[1] The definition of offensive language has been taken up later by *Founta*.

Incomplete Information for Definitions. The *TRAC* definitions of 'overt aggression' and 'covert aggression' are also very generic; 'covert aggression' is simply defined as negation of 'overt aggression', which does not provide enough information about the class. For the toxicity dataset provided in *Kaggle*, there are no specific definitions of the categories. I assume that they are the same as the ones used in the context of the Perspective API.

The next step is label standardization, which must bear in mind which definitions are explicit or ambiguous, distinct or similar, or incomplete.

4.1.3 Label standardization between datasets

To be able to compare the different categories across the datasets, I standardize the label categories by assigning the same labels to the equivalent categories in the different datasets. The standardization relies on analysing dataset categories, properties, definitions, and data collections presented in Section 5.1.1 and the observations on these definitions in Section 4.1.2. Table 4.1 shows the outcome of the standardization.

Let me illustrate, in what follows, the steps of this standardization procedure. For instance, in the case of the *W&H* dataset, the 'sexism' and 'racism' categories are considered both as separate categories, but also as subcategories of 'hate speech'. Hence, a new category ('hate speech') is added to this dataset to increase compatibility with other annotations. Furthermore, the 'sexism' category in this dataset is assumed to be equivalent to the 'misogynous' category of the *Ami* dataset since in the literature no clear distinction between these two categories is provided. The resulting standardized cross-dataset label is called 'misogyny-sexism'. For the *Davidson* dataset, a new category 'toxicity' is introduced that subsumes the union of its 'hate speech' and 'offensive' categories. 'Toxicity' is an umbrella term that aims to capture general offense and different types of 'aggression' (Kolhatkar et al., 2020). *Ami's* 'misogynous' category is assumed to be equivalent to the 'sexism' category in *W&H*'s dataset, as already mentioned. The *TRAC* dataset contains the categories 'overt aggression' and 'covert aggression', which are merged into a new category 'aggression'. It would have been possible to also convert them into a general category such as 'toxicity', however, this was not done for *TRAC*, as it aims to identify subtler aggression, which is a dimension not mentioned in the *Kaggle* dataset. Regarding the *Hateval* dataset, its 'aggression' category covers a specific type of aggression as it is a subset of 'hate speech'. In this case, the two categories are not merged into 'toxicity', as *Hateval* aims to classify only 'hate speech', and considers 'aggression' only when it happens in the context of 'hate speech'. Moreover, it is considered 'aggressive hate speech' and not equivalent to the 'aggression' category in *TRAC* dataset. For the *Kaggle* dataset, the original labels are kept, as the dataset is already annotated in a multiclass manner. For *Stormfront*, the original category is also kept since the authors use a 'hate speech' definition that focuses on the target characteristics of this type of communication (e.g., gender and age). This is similar to previous definitions found in the literature, and the aim is to test whether the different 'hate speech' annotated datasets generalize within themselves. In the *Offenseval* dataset, only 'offense' is annotated as a general category that is meant to cover all types of 'offensive' speech. Therefore, it is converted into 'toxicity', such that it becomes comparable to the equally general terms found in the *Davidson* and *Kaggle* datasets (again, using the criteria defined in Fortuna et al. (2020)). Finally, the *Founta* dataset is annotated with 'hateful', 'abusive', 'spam', and 'normal' labels. 'spam' is considered to fall into the category 'normal', as this study is interested in abusive speech only. In the standardized category scheme, both ('normal' and 'spam') are marked as 'none'. In contrast, the original 'abusive' label is kept. Although the authors mention that 'abusive' significantly correlates with 'aggression' and 'offensive' categories, it is the most popular label among the three, the most central in *Founta*, and it is the label that the authors preferred for their dataset. Furthermore, it seems that this category is not equivalent to 'aggression' from *TRAC*, as it does not include the covert and overt dimensions. In both cases, their conversion into the 'offensive' category of *Davidson* was not considered, since this conversion would lose information. Finally, the category 'hateful' is converted into 'hate speech', to be in line with the definition in the literature. The resulting conversion is presented in Table 4.1.

Overall, the standardization procedure can be described as:

- Keep the original labels when possible.
- Group categories if this increases the compatibility between datasets.
- Rename labels to increase dataset compatibility.
- Rename labels when similar names are used for different phenomena.

4.1.4 Cluster Categories

After establishing a concept-driven standardization of the classes in the various datasets, I perform additional analysis to obtain more information on the classes and datasets. The categories are compared across the datasets with respect to both their similarity to the other categories and their homogeneity, i.e., variation of the samples of one single category. A non-supervised approach is followed.

General Procedure

For clustering, each category is represented as a centroid vector using Fast Text (Bojanowski et al., 2017) and pretrained word embeddings trained on Wikipedia (Mikolov et al., 2018). This approach is followed because the majority of the datasets contain short texts generated in social networks and Fast Text along with pretrained word embeddings has been providing good results in different works applied to the automatic detection of hate speech and related concepts with similar data; see, e.g., (Santucci et al., 2018; Fortuna and Nunes, 2019).

The process I use to compute the aforementioned message centroids is as follows:

Table 4.1: Used standardized categories (for convenience, in the text 'misogyny-sexism' is referred as 'sexism').

dataset	**original category**	**standardized category**
W&H	sexism	misogyny-sexism
	racism	racism
	sexism or racism	hate speech
Davidson	hate speech	hate speech
	offensive	offensive
	hate speech or offensive	toxicity
Ami	misogynous	misogyny-sexism
TRAC	covert aggression	covert aggression
	overt aggression	overt aggression
	overt or covert aggression	aggression
Hateval	hate speech	hate speech
	aggression	aggressive hate speech
Kaggle	threat	threat
	identity hate	hate speech
	severe toxic	severe toxic
	insult	insult
	obscene	obscene
	toxic	toxicity
Stormfront	hate speech	hate speech
Offenseval	offensive	toxicity
Founta	hateful	hate speech
	abusive	abusive
	spam	none
	hateful or abusive	toxicity

- Pre-process the messages by lowercasing all words, removing IPs, Twitter elements such as hashtags, usernames, and stop words using NLTK.

- Train word embeddings using FastText and the 300-dimension English Wikipedia pretrained embeddings.

- Extract the centroid of the message by averaging the word embeddings of each of its sentences.

Inter-dataset class similarity

In this experiment, the goal is to compare the different categories across the annotated datasets in terms of their semantic similarity. For this, in addition to the previous procedure a further step is taken:

- Compute the average of every message centroid that belongs to each category, obtaining the centroid of each category.

After obtaining the category centroids, a Principal Component Analysis (PCA) (Pearson, 1901) is performed to obtain a 2D representation and thus be able to plot the centroids. The result of this process can be seen in Figure 4.1.

To complement the visualization of the category centroids, the distances between each pair of categories and the standardized category labels is computed (see Table 4.1) to get a better grasp on how similar these categories actually are. The cosine distance metric provides the formula to compute the distances.

The analysis of the PCA plot and the inter-class distance analysis are presented in Subsections 4.1.4 and 4.1.4. Both analyses are distinct and complementary: PCA represents the distance between classes when considering the feature reduction to two orthogonal dimensions, while the inter-class distance compares all messages of the corresponding classes. In other words, inter-class distance is a metric that measures the similarity between two classes in terms of how much the messages of the two vary.

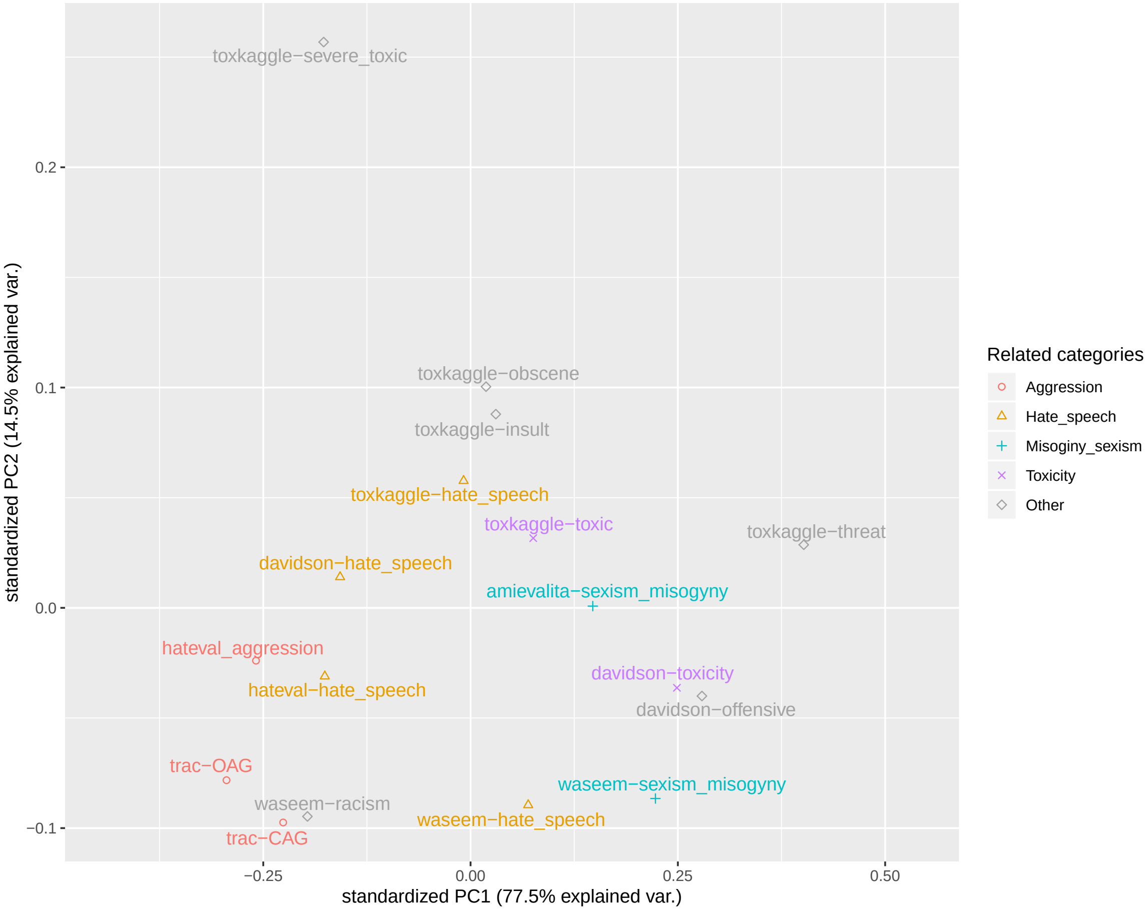


Figure 4.1: PCA results

Intra-dataset class homogeneity

In this experiment, different categories are compared across the annotated datasets with respect to their internal homogeneity. For this purpose, in addition to the procedure described in Subsection 4.1.4, the following steps are applied:

- Compute the distance between all the messages from the same category by using the cosine similarity.
- Average the distances in order to estimate the homogeneity of a category.

The analysis of the intra-class distances is presented in Subsection 4.1.4.

Results and Analysis of the Inter and Intra-dataset Class Homogeneity

Centroid Visualization To create the graph shown in Figure 4.1, the two first principal components of the category centroids are selected. The goal is to see how the different categories relate to each other, hence the plot displays a different color for possibly related categories. The results seem to be coherent with what is expected as there are clear similarities between classes that represent similar categories. For instance, the 'aggression' related categories (in red) tend to be grouped together and intersected with hate speech, as expected for the 'Hateval-aggression' category. The hate speech categories (in yellow) also appear close in space. From these categories, 'Davidson-hate speech' and 'Hateval-hate speech' are the closest, while 'W&H-hate speech' is at same time close to 'hate speech' but also between 'W&H-sexism' and 'W&H-racism' – again as expected.

The 'Kaggle-identity hate' category appears close to 'Davidson-hate speech', but, in this case, closer to other "Kaggle" categories ('toxic', 'insult' and 'obscene'). This is probably due to the multiclass property of the 'Kaggle' dataset, where the same message can have different labels.

This property, and also the fact that this is the only dataset collected from Wikipedia comments may justify why the 'Kaggle' dataset categories are mapped together in the upper part of the figure and are more difficult to compare to the categories of the other datasets. However, at same time 'Kaggle-severe-toxic' is far from the same dataset categories, including 'Kaggle-toxic'.

Apart from conforming an expected degree of similarity between specific categories of the different datasets, the PCA plot allows us to gather new insights about the data. For instance, the general categories 'toxicity' from Kaggle ('Kaggle-toxic') and 'aggression' from TRAC ('Trac-CAG' or 'Trac-OAG') do not appear close, despite the fact that both 'toxicity' and 'aggression' are defined as general umbrella terms for offensive, toxic or abusive online behavior. In contrast, 'Kaggle-toxic' and 'Davidson-toxicity', which are assigned during the category standardization the label 'toxicity', appear closer in the plot. Additionally, between these two categories, 'Ami-sexism-misoginy' is situated, indicating that 'sexism' can be one of the main types of toxicity in those datasets.

Also, the 'misoginy-sexism' related categories ('Ami-sexism-misoginy' and 'W&H-sexism') seem close, but 'Davidson-offensive' categories seem more similar to them. Another interesting observation is that the category 'W&H-racism' seems to be very close to both TRAC dataset categories, indicating that racism can be more frequent than other categories in the TRAC dataset.

Inter Dataset Class Distance To further analyze how similar or dissimilar the categories across the datasets are, the distances between class centroids are inspected. Table 4.2 shows each category of each dataset (in bold as header) and the top 5 most similar categories (below the header).[2] As expected, these results are aligned with the PCA. For the hate speech related categories, 'Davidson-hate speech' is close to 'Kaggle-identity

[2] The full table with the distance values for each pair of categories can be found at `https://docs.google.com/spreadsheets/d/1mkSTmuO8cc8tUbAEq68J_el39hyx6uvEWo1xPFGMRvg/edit?usp=sharing`.

hate' and Hateval's 'hate speech', but farther from W&H's 'hate speech', 'racism', 'sexism' and 'Ami-misogyny'. This seems to indicate that there are several different representations of the notions of 'hate speech' and its subtypes.

'Ami-sexism-misogyny' appears to be close to Davidson's 'offensive' and Kaggle's 'toxicity', but it is also not so far away from W&H's 'sexism'. On the other side, 'W&H-sexism' is also closer to 'Kaggle-toxic' than to 'Ami-sexism-misogyny'. This may indicate that the Kaggle 'toxicity' category contains sexist messages that are more similar to the 'W&H-sexist' messages. Nevertheless, it is surprising that W&H's 'sexism' category appears more similar to 'toxicity' than to 'Ami-misogyny'.

Regarding the Kaggle categories, its 'identity hate' is close to its 'insult', 'toxic' and 'obscene' categories, and more distant to the hate speech categories from the other datasets (i.e., 'Davidson-hate speech' and 'Hateval-hate speech', or 'Ami-sexism-misogyny'). This indicates that in this dataset the category notions are very interdependent. Even more obvious is the overlap between Kaggle's 'insult' and 'obscene', which are very close to each other and largely share the distances to the other categories. Indeed, the distinction between both is not clear.

Kaggle's 'severe toxic' is closer to all the other Kaggle's dataset categories, but the reverse does not apply. Thus, 'Kaggle-toxic' is closer to 'Kaggle-insult', followed by 'Kaggle-identity hate' and 'Ami-sexism-misogyny', and very far from 'severe toxic', which is quite unexpected, since their labels suggest that the main difference between these two categories is the intensity of the expressed toxicity.

Intra-Category Homogeneity Figures 4.2 and 4.3 display the homogeneity of the individual categories for each dataset in terms of the average distance between their messages. The more homogeneous a category is, the smaller the value in the plot. The most homogeneous category is 'W&H-racism'. Specific types of harmful content such as 'racist', 'misogynous' and 'threats' appear to be quite homogeneous, which indicates that these categories are well-defined, and its messages are clearly identifiable. On the other hand, hate speech presents various homogene-

trac-cag	**trac-oag**	**davidson-toxicity**	**davidson-hate_speech**
trac-oag	trac-cag	davidson-offensive	toxkaggle-identity_hate
waseem-racism	hateval-aggression	amievalita-misogynous	hateval-hate_speech
hateval-aggression	waseem-racism	waseem-sexism	hateval-aggression
hateval-hate_speech	hateval-hate_speech	waseem_hate_speech	toxkaggle-toxic
waseem_hate_speech	waseem_hate_speech	toxkaggle-toxic	toxkaggle-obscene
davidson-offensive	**amievalita-misogynous**	**hateval-aggression**	**hateval-hate_speech**
davidson-toxicity	davidson-toxicity	hateval-hate_speech	hateval-aggression
amievalita-misogynous	davidson-offensive	waseem-racism	waseem-racism
waseem-sexism	toxkaggle-toxic	trac-oag	trac-cag
waseem_hate_speech	waseem-sexism	trac-cag	waseem_hate_speech
toxkaggle-toxic	toxkaggle-insult	davidson-hate_speech	trac-oag
toxkaggle-identity_hate	**toxkaggle-insult**	**toxkaggle-obscene**	**toxkaggle-threat**
toxkaggle-insult	toxkaggle-obscene	toxkaggle-insult	toxkaggle-toxic
toxkaggle-toxic	toxkaggle-toxic	toxkaggle-toxic	amievalita-misogynous
toxkaggle-obscene	toxkaggle-identity_hate	toxkaggle-identity_hate	toxkaggle-insult
davidson-hate_speech	amievalita-misogynous	amievalita-misogynous	waseem-sexism
hateval-hate_speech	hateval-hate_speech	hateval-hate_speech	toxkaggle-obscene
toxkaggle-severe_toxic	**toxkaggle-toxic**	**waseem_hate_speech**	**waseem-racism**
toxkaggle-obscene	toxkaggle-insult	waseem-sexism	hateval-aggression
toxkaggle-insult	toxkaggle-identity_hate	hateval-hate_speech	hateval-hate_speech
toxkaggle-identity_hate	toxkaggle-obscene	toxkaggle-toxic	trac-cag
toxkaggle-toxic	amievalita-misogynous	waseem-racism	trac-oag
davidson-hate_speech	waseem_hate_speech	amievalita-misogynous	waseem_hate_speech
waseem_sexism			
waseem_hate_speech			
toxkaggle-toxic			
amievalita-misogynous			
davidson-toxicity			
davidson-offensive			

Table 4.2: Top 5 most similar to each label

ity scores: 'W&H-hate _speech' is quite homogeneous, since it is composed of racist and misogynous messages while Davidson's hate speech instances are very heterogeneous, which is coherent with its definition, where messages that express hatred toward any target group are considered (see Tables 2.3 and 2.4). In the light of the findings of Arango et al. (2019), a paper that discusses how 'W&H' hateful data originates from a small number of users, it is also unsurprising that the categories in this dataset are more homogeneous.

The assessment of the number of messages per category shows that the homogeneity is not affected by it. Furthermore, homogeneity does not depend on the dataset, neither on the platform used for data collection.

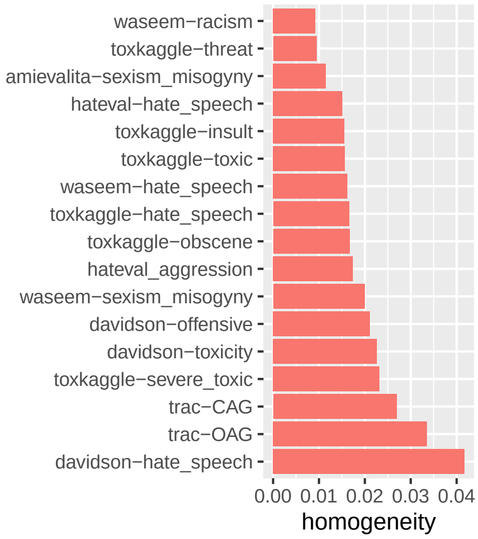


Figure 4.2: Class homogeneity

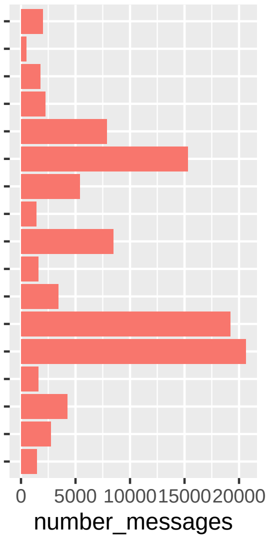


Figure 4.3: Number of messages per class.

4.2 Chapter Discussion

The analysis conducted in this chapter shows that the quality of the datasets in terms of their annotation should be improved. The results of the analysis also suggest that the intra and inter-dataset coherence of the annotation should equally be improved. The following guidelines are intended to help addressing both problems:

- Avoid creating new categories to refer to concepts already present in the literature. In the case a new category is identified, provide clear examples and justification why a new category is needed.

- In case of assessing that a new concept is needed, position it in the map of existing categories when annotating datasets, for instance, by following a similar method to the one provided in this chapter. Different publicly available English datasets containing abuse language are annotated with respect to their similarity and compatibility. In the future, studies on other, new, datasets should conduct the same type of analysis. Different feature extraction procedures will allow to reveal similarity to other dimensions such as topics.

- Provide detailed information on the sampling procedure; for instance, the used data source (e.g., Twitter), the groups that are targeted (e.g., women), the context to which the data refer to (e.g., comments on news about politics, or sports), the time and location for the data. Discuss geographical, cultural and socio-political values of concepts and datasets to better characterize the data.

- Provide detailed information on the class balancing procedure. The proportion between offensive, toxicity, abuse or hate messages can vary across different datasets.

- Follow a systemic discrimination theory for hate speech, which suggests to acknowledge that defining and identifying hate speech encompasses the privilege of deciding which minorities and patterns of speech qualify for hate speech utterances (Gelber, 2021). Such selection mechanisms are not made clear nor discussed in the context of the available datasets. The majority of the works considers general hate speech, or even broader concepts such as 'offense', leaving to the annotators the decision to define which groups are targeted and which expressions qualify as hate speech. Only *W&H* and *Ami* refer to more specific hate speech classes ('sexism' and 'racism' in the first and 'misogyny' in the second).

- Document the geographic and cultural features of a dataset, such as, the geographic and cultural background in which the messages were generated, the background of the annotators and the background underlying the defined hate speech concepts to annotate. These features are largely ignored in the available datasets, which makes it impossible to assess whether the datasets under inspection imply the same context, thus affecting our ability to accurately perform categories standardization.

4.3 Chapter Summary

The present chapter aimed at clarifying categories and comparing a selection of publicly available datasets – a question that has been already raised in the literature. It contributed to the understanding of how definitions in the different datasets relate and in which cases it is possible to make datasets compatible.

In conformity with the ideas discussed in Chapter 3, it has been found that indeed a plethora of definitions for hate speech exists and different datasets follow different terminologies. However, the selection and construction of definitions is not always clear. Opting for one definition or another is a power decision that has an effect on the annotated labels and the capabilities of the resulting models. This is an aspect that should be better documented and receive more attention in the future.

Because there was no corresponding information on the datasets, the analysis in the present chapter does not take into account geographical, cultural and socio-political features when looking at the hate speech definition and dataset construction across different datasets. Nevertheless, these are important factors that need to be taken into account.

In the past, studies of individual datasets prevailed in hate speech research. Now, with the standardization of different categories at hand, multidataset research is possible and may provide a wealth of additional insights.

The analysis presented in this chapter is a necessary step to establish a framework for dataset compatibility, which will be used in the next chapter to explore the possibility of testing model generalization via cross-dataset experiments and to better investigate classification conventions for hate speech detection.

Chapter 5

GAINING INSIGHTS ON MODEL EVALUATION

In Section 3.1.3, I discuss the over-estimation of the high performance figures obtained during the evaluation of the models for hate speech detection. Not only have models been promoted as having superior performance to reality, but they have also been presented as having superior generalization potential. Current conventions for testing model generalization foresee the evaluation of the model with different data than used for training, but still with data from the same dataset as the training data subset, which permits the model to be aware of data that possess some expected characteristics such as class distribution.

In this chapter, I discuss how another protocol for testing of hate speech detection models helps to get a better understanding of the model's capabilities and generalizability. I focus on studying model evaluation and the lack of generalizability of models trained to detect hate speech by using cross-dataset experiments, i.e., experiments in which training and testing data are derived from distinct sample efforts. If hate speech were a universal concept, a single classifier would be applicable to different contexts, even with samples that are distinct from the training data. When employing classifiers in cross-dataset experiments, however, generalization has not been as straightforward to achieve. Existing research on cross-dataset experiments for hate speech related classifiers has previously been conducted; see, e.g., Waseem et al. (2018); Karan and Šnajder (2018); Arango et al. (2019, 2020); Swamy et al. (2019); Salminen et al. (2020); Chandrasekharan et al. (2017). Nevertheless, such studies and efforts can be extended in order to better address the specific problem of hate speech. Previous cross-dataset studies are limited since they consider only a limited number of datasets (and thus a limited number of categories), as, e.g.,Waseem et al. (2018) considers two datasets only, or they merge all categories into one positive[1] category (e.g., 'abusive'), which is then contrasted to a negative category (e.g., 'not abusive'), as, e.g., Karan and Šnajder (2018) and Swamy et al. (2019). Both the limitation in scope and the fusion of the original categories of the different datasets into one generic category impede a conclusive answer on the generalization potential of models classifying abusive language – and most importantly for this thesis, hate speech. To address this problem, I analyze the cross-dataset performance of two state-of-the-art models, BERT (Devlin et al., 2019) and ALBERT (Lan et al., 2020), and two baselines, fastText (Bojanowski et al., 2017), and SVM (Pedregosa et al., 2020), trained on nine of the most common abusive language datasets in English. To be able to compare the performance of such models across datasets, the original dataset categories are standardized using the conversion schema proposed in Chapter 4. To better understand the generalization potential of the classifiers, the performance of 450 BERT, 450 ALBERT, 450 fastText, and 348 SVM binary abusive language classifiers (1698 in total) is accessed based on different features, using a Random Forest model. I also provide an additional experiment using the Perspective API, to understand generalization of general purpose toxicity models when applied to more particular subtypes of toxicity such as hate speech.

The objectives of this chapter include refining model testing conventions and comprehending the lack of generalizability of models for hate speech (Goals 5 and 6 of the thesis). The bulk of the reported experiments and findings has been published in Fortuna et al. (2020) and Fortuna et al.

[1] With "positive category" I refer to the class to detect, as is traditionally done in ML, e.g Sebastiani (2002).

(2021).

5.1 Evaluating Model Generalizability Across Hate Speech Related Datasets

In this section, I describe a series of experiments that enables me to assess the generalizability of a model. In the first experiment, generalizability is tested using the typical methods of partitioning the sample into train-validation-test sets and using conventional metrics in an intra-dataset scenario. This means that the model is evaluated using data received concurrently with the training set and likely sharing other attributes such as vocabulary, users, or topics. In a further step, evaluation is conducted via a cross-dataset experiment, in which models are trained on one dataset but assessed on data from another.

Before I begin with the presentation of the experiments, let me describe the datasets and employed models.

5.1.1 Datasets

Nine publicly available datasets are used from the list in Table 2.2 that cover different offensive, abusive language and hate speech related categories: *W&H*, *Davidson*, *Ami*, *Stormfront*, *TRAC*, *Kaggle*, *Hateval*, *Offenseval*, and *Founta*.

From most of the datasets, only the training partitions of the datasets are considered since their test sets are not always available and in cases they are, the splits between the training and test sets vary. The training sets are split randomly into 70% for training and 30% for testing the models. The exceptions are *Hateval* and *Offenseval*, of which both the training and the test sets are used in their original 70%–30% split. This training–test set division per dataset (see Table 5.1) is kept all over the experiments, hence also for all the standardized categories of a given dataset.

To obtain an objective picture of the generalization potential of the models across different datasets, the procedure for category standardization presented in Chapter 4 is followed. The obained class frequencies are presented in Table 5.1.

5.1.2 Models

For the experiments, BERT is selected as it is a very relevant transformer model and ALBERT because it has been reported to outperform BERT (Lan et al., 2020). fastText and SVM are used as baselines. The macro averaged F1 score is used for evaluation of all the models. For BERT, ALBERT and FastText, off-the-shelf models are used.[2] For the SVM experiments, 10-fold cross-validation is applied instead of a 70%/30% split.[3] The training of BERT and ALBERT is carried out on a TPU in COLAB with Tensorflow 1.15.[4] In the case of BERT and ALBERT, L refers to the number of layers or Transformer blocks, H to the hidden size, and A to the number of self-attention heads.

In what follows, the setup of each model is presented.

BERT BERT$_{\text{LARGE}}$ (L=24, H=1024, A=16) is used with 340M parameters in total, which outperforms BERT$_{\text{BASE}}$ across all tasks, especially those with very little training data (Devlin et al., 2019), as is the case with some of the used datasets. Additionally, a batch size of 32 and fine-tune for 3 epochs over the data of all datasets are used. The dropout probability is set to 0.1 for all layers; the Adam optimizer is used with a learning rate of 2e-5.

ALBERT ALBERT$_{\text{XXLARGE}}$ (L=12, H=4096, A=64) model is used with about 70% of BERT$_{\text{LARGE}}$'s parameters for an available trained ALBERT model (Lan et al., 2020), fine-tuned to all nine datasets as described in

[2] https://fasttext.cc/docs/en/english-vectors.html https://github.com/google-research/bert https://github.com/google-research/albert

[3] For this purpose, the training and test sets of *Offenseval* and *Hateval* are merged.

[4] https://github.com/paulafortuna/IP-M_abusive_models_generalize

Table 5.1: Dataset and respective standardized category (st. category), total number of instances for training (train total N), total number of instances for test (test total N), total number of positive instances for training (train total pos), and percentage of positive instances in the training set (train perc positive).

dataset	**st. category**	**train total N**	**test total N**	**train total pos**	**train perc positive**
W&H	sexism	11835	5073	2407	0.20
W&H	racism	11835	5073	1377	0.12
W&H	hate speech	11835	5073	3784	0.32
Davidson	hate speech	17348	7435	975	0.06
Davidson	offense	17348	7435	13517	0.78
Davidson	toxicity	17348	7435	14492	0.84
Ami	sexism	2800	1200	1249	0.45
TRAC	covert aggression	12000	3000	4240	0.35
TRAC	overt aggression	12000	3000	2708	0.23
TRAC	ov cov aggression	12000	3000	6948	0.58
Hateval	hate speech	9000	1000	3783	0.42
Hateval	aggressive hs	9000	1000	1559	0.17
Kaggle	toxicity	111699	47872	10856	0.10
Kaggle	hate speech	111699	47872	977	0.01
Kaggle	severe toxicity	111699	47872	1107	0.01
Kaggle	insult	111699	47872	5593	0.05
Kaggle	obscene	111699	47872	6008	0.05
Kaggle	threat	111699	47872	336	0.00
Stormfront	hate speech	3501	1500	705	0.20
Offenseval	toxicity	13240	319	4400	0.33
Founta	hate speech	64366	27587	2885	0.05
Founta	abusive	64366	27587	14463	0.23
Founta	toxicity	64366	27587	17348	0.27

Lan et al. (2020). A batch size of 32 is used, the dropout probability is set to 0.1 for all layers and the Adam optimizer is used with a learning rate of 1e-5.

FastText FastText is similar to the CBOW model as, unlike standard bag-of-words model, it uses continuous distributed representation of the context (Mikolov et al., 2013). The model is ran in its version 0.9.2 with 300 dimension-FastText pretrained vectors (Mikolov et al., 2018), Skip-gram Hierarchical softmax loss function, learning rate of 1.0, considering 1 as minimal number of word occurrences, bi-grams, and 25 epochs.

BOW + SVM In the BOW+SVM experiments, the Scikit Learn models are used (Pedregosa et al., 2011).[5] For the Bag-Of-Words (BOW) extraction, stopwords are removed and considered only words with a frequency $\geq 1\%$. For SVM classification, I use most of its default parameters, except for the kernel, which is set to the linear kernel. Due to the time complexity of the parameter extraction and training procedures, I use SVM with bagging. The time complexity, paired with the size of the dataset, also forces to exclude the *Kaggle* dataset from this classification task.

5.1.3 Intra-Dataset Model Evaluation

In this experiment, I create binary classification models for BERT, ALBERT, fastText and SVM. Figures 5.1 and 5.2 display the results of BERT / ALBERT / fastText / SVM for intra-dataset classification in terms of the macro averaged F1 score, grouped by standardized categories (Figure 5.1) and datasets (Figure 5.2).

5.1.4 Cross-Dataset Model Evaluation

The obtained intra-dataset models are tested in a second experiment in a cross-dataset scenario. More precisely, each model, trained on a spe-

[5] For BOW, I use the *CountVectorizer* class, for SVM the SVC class and for bagging the *BaggingClassifier* class.

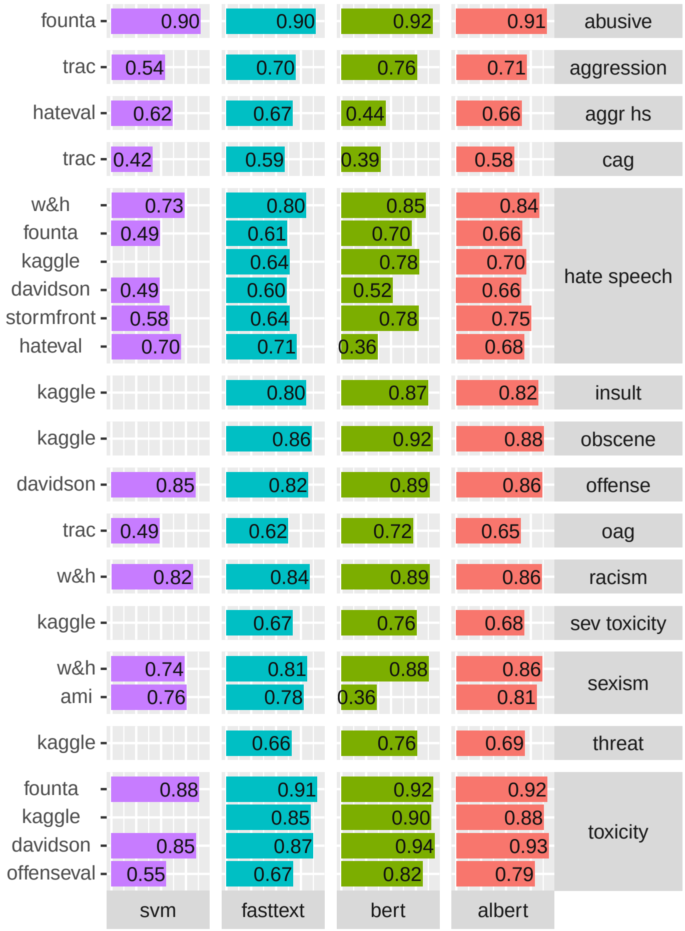


Figure 5.1: Macro F1 scores, by standardized categories (‘cag’: covert aggression; ‘oag’: overt aggression, ‘sev toxicity’: severe toxicity, ‘aggr hs’: aggressive hate speech).

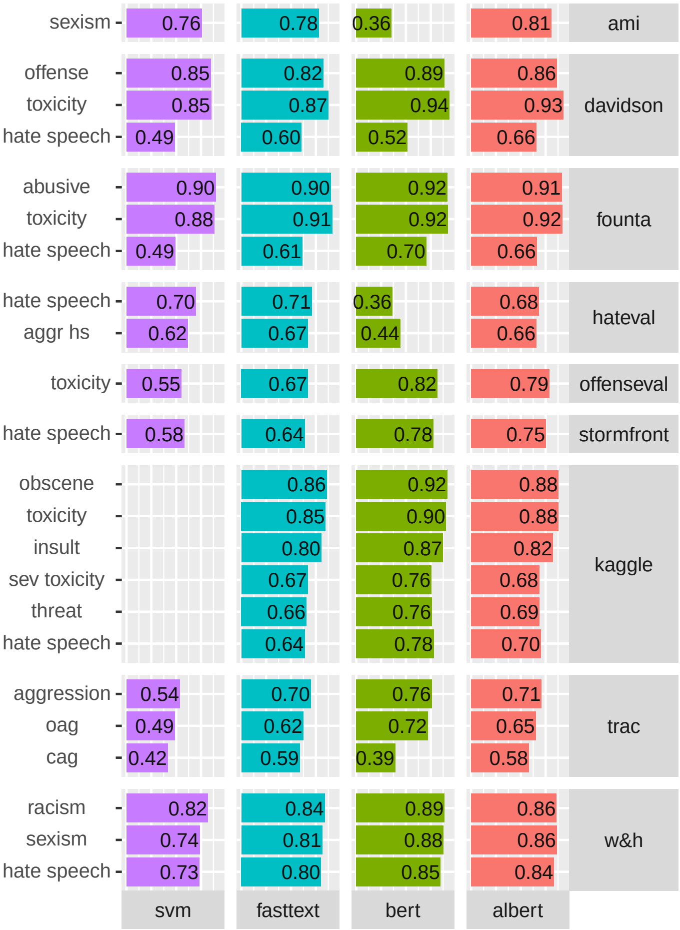


Figure 5.2: Macro F1 scores, by datasets.

cific dataset, is tested against all the remaining datasets and corresponding standardized categories. The same training and test set divisions apply to intra-dataset and cross-dataset experiments. As already before, the macro averaged F1 measure is used for evaluation. The results of the experiment on inter-(or cross-)dataset classification are displayed in Table 5.2.[6]

Due to space constraints, only the results with F1 score ≥ 0.60 are displayed.[7] It is assumed that there is a better cross-dataset generalization if at least one of the four algorithms (BERT, ALBERT, fastText, or SVM) achieves with its model an F1 score of ≥ 0.70.

5.1.5 Model Performance Classification

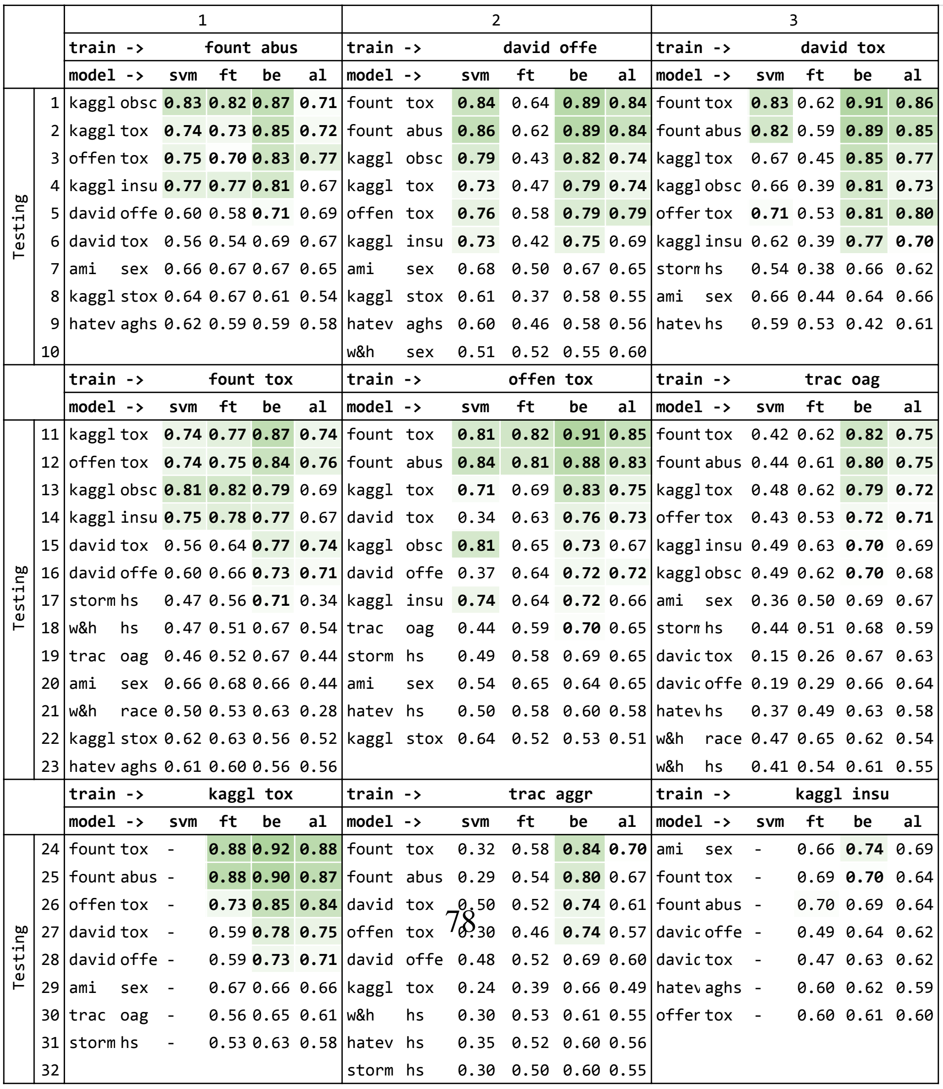
To systematically study which model and dataset features lead to a better generalization in abusive language-related models, an experiment is carried out on the relation between the performance figures obtained when applying BERT, ALBERT, fastText, and SVM and 16 prominent features of the models and datasets considered in the literature as good generalization predictors; see Table 5.4. For this purpose, the 1698 binary BERT/ALBERT/fastText/SVM models (450 of each for BERT/ALBERT/fastText and 348 for SVM) are grouped into models that generalize better (those with an F1 score ≥ 0.70; 136 in total) and models that generalize worse (those with an F1 score < 0.70; 1562 in total). The goal is to train a classifier on the above 16 features to predict whether a model belongs to the better generalizing models or worse generalizing models. As classifier, Random Forest is used with 50 estimators (Pedregosa et al., 2011) with 5 Fold cross-validation. Random Forest is a general-purpose classifier with weak statistical assumptions which makes it a good option for this experiment. To rank the different features used for classification, the permutation feature importance algorithm is applied.[8] It directly mea-

[6] The shades in the table cells reflect the F1 score: from white (F1 $\leq$0.69) to green (F1 = 1.0).

[7] Note that in what follows, the dataset name abbreviations are used as introduced in Table 5.2

[8] `https://explained.ai/rf-importance/index.html`.

Table 5.2: Model generalization evaluation of BERT (be), ALBERT (al), fastText (ft) and SVM (svm) in terms of macro F1 score. The second row from the top in bold indicates the dataset and standardized category used for training. The remaining rows (second column) of the test data (*Offenseval* (‘offen’), *Davidson* (‘david’), *Founta* (‘fount’), *Kaggle* (‘kaggl’), *Hateval* (‘hatev’) and categories toxicity (‘tox’), obscene (‘obsc’), insult (‘insu’), aggression (‘aggr’), offensive (‘offe’), overt aggression (‘oag’), covert aggression (‘cag’), abusive (‘abus’), hate speech (‘hs’), sexism (‘sex’), racism (‘race’), severe toxicity (‘stox’), aggressive hate speech (‘aghs’)).



		1						2						3					
		train ->		**fount**	**abus**			**train ->**			**david**	**offe**		**train ->**			**david**	**tox**	
		model ->		**svm**	**ft**	**be**	**al**	**model ->**		**svm**	**ft**	**be**	**al**	**model ->**		**svm**	**ft**	**be**	**al**
Testing	1	kaggl	obsc	**0.83**	**0.82**	**0.87**	**0.71**	fount	tox	**0.84**	0.64	**0.89**	**0.84**	fount	tox	**0.83**	0.62	**0.91**	**0.86**
	2	kaggl	tox	**0.74**	**0.73**	**0.85**	**0.72**	fount	abus	**0.86**	0.62	**0.89**	**0.84**	fount	abus	**0.82**	0.59	**0.89**	**0.85**
	3	offen	tox	**0.75**	**0.70**	**0.83**	**0.77**	kaggl	obsc	**0.79**	0.43	**0.82**	**0.74**	kaggl	tox	0.67	0.45	**0.85**	**0.77**
	4	kaggl	insu	**0.77**	**0.77**	**0.81**	0.67	kaggl	tox	**0.73**	0.47	**0.79**	**0.74**	kaggl	obsc	0.66	0.39	**0.81**	**0.73**
	5	david	offe	0.60	0.58	**0.71**	0.69	offen	tox	**0.76**	0.58	**0.79**	**0.79**	offer	tox	**0.71**	0.53	**0.81**	**0.80**
	6	david	tox	0.56	0.54	0.69	0.67	kaggl	insu	**0.73**	0.42	**0.75**	0.69	kaggl	insu	0.62	0.39	**0.77**	**0.70**
	7	ami	sex	0.66	0.67	0.67	0.65	ami	sex	0.68	0.50	0.67	0.65	storm	hs	0.54	0.38	0.66	0.62
	8	kaggl	stox	0.64	0.67	0.61	0.54	kaggl	stox	0.61	0.37	0.58	0.55	ami	sex	0.66	0.44	0.64	0.66
	9	hatev	aghs	0.62	0.59	0.59	0.58	hatev	aghs	0.60	0.46	0.58	0.56	hatev	hs	0.59	0.53	0.42	0.61
	10							w&h	sex	0.51	0.52	0.55	0.60						
		train ->		**fount**	**tox**			**train ->**			**offen**	**tox**		**train ->**			**trac**	**oag**	
		model ->		**svm**	**ft**	**be**	**al**	**model ->**		**svm**	**ft**	**be**	**al**	**model ->**		**svm**	**ft**	**be**	**al**
Testing	11	kaggl	tox	**0.74**	**0.77**	**0.87**	**0.74**	fount	tox	**0.81**	**0.82**	**0.91**	**0.85**	fount	tox	0.42	0.62	**0.82**	**0.75**
	12	offen	tox	**0.74**	**0.75**	**0.84**	**0.76**	fount	abus	**0.84**	**0.81**	**0.88**	**0.83**	fount	abus	0.44	0.61	**0.80**	**0.75**
	13	kaggl	obsc	**0.81**	**0.82**	**0.79**	0.69	kaggl	tox	**0.71**	0.69	**0.83**	**0.75**	kaggl	tox	0.48	0.62	**0.79**	**0.72**
	14	kaggl	insu	**0.75**	**0.78**	**0.77**	0.67	david	tox	0.34	0.63	**0.76**	**0.73**	offer	tox	0.43	0.53	**0.72**	**0.71**
	15	david	tox	0.56	0.64	**0.77**	**0.74**	kaggl	obsc	**0.81**	0.65	**0.73**	0.67	kaggl	insu	0.49	0.63	**0.70**	0.69
	16	david	offe	0.60	0.66	**0.73**	**0.71**	david	offe	0.37	0.64	**0.72**	**0.72**	kaggl	obsc	0.49	0.62	**0.70**	0.68
	17	storm	hs	0.47	0.56	**0.71**	0.34	kaggl	insu	**0.74**	0.64	**0.72**	0.66	ami	sex	0.36	0.50	0.69	0.67
	18	w&h	hs	0.47	0.51	0.67	0.54	trac	oag	0.44	0.59	**0.70**	0.65	storm	hs	0.44	0.51	0.68	0.59
	19	trac	oag	0.46	0.52	0.67	0.44	storm	hs	0.49	0.58	0.69	0.65	davic	tox	0.15	0.26	0.67	0.63
	20	ami	sex	0.66	0.68	0.66	0.44	ami	sex	0.54	0.65	0.64	0.65	davic	offe	0.19	0.29	0.66	0.64
	21	w&h	race	0.50	0.53	0.63	0.28	hatev	hs	0.50	0.58	0.60	0.58	hatev	hs	0.37	0.49	0.63	0.58
	22	kaggl	stox	0.62	0.63	0.56	0.52	kaggl	stox	0.64	0.52	0.53	0.51	w&h	race	0.47	0.65	0.62	0.54
	23	hatev	aghs	0.61	0.60	0.56	0.56							w&h	hs	0.41	0.54	0.61	0.55
		train ->			**kaggl**	**tox**		**train ->**			**trac**	**aggr**		**train ->**			**kaggl**	**insu**	
		model ->		**svm**	**ft**	**be**	**al**	**model ->**		**svm**	**ft**	**be**	**al**	**model ->**		**svm**	**ft**	**be**	**al**
Testing	24	fount	tox	-	**0.88**	**0.92**	**0.88**	fount	tox	0.32	0.58	**0.84**	**0.70**	ami	sex	-	0.66	**0.74**	0.69
	25	fount	abus	-	**0.88**	**0.90**	**0.87**	fount	abus	0.29	0.54	**0.80**	0.67	fount	tox	-	0.69	**0.70**	0.64
	26	offen	tox	-	**0.73**	**0.85**	**0.84**	david	tox	0.50	0.52	**0.74**	0.61	fount	abus	-	0.70	0.69	0.64
	27	david	tox	-	0.59	**0.78**	**0.75**	offen	tox	0.30	0.46	**0.74**	0.57	davic	offe	-	0.49	0.64	0.62
	28	david	offe	-	0.59	**0.73**	**0.71**	david	offe	0.48	0.52	0.69	0.60	davic	tox	-	0.47	0.63	0.62
	29	ami	sex	-	0.67	0.66	0.66	kaggl	tox	0.24	0.39	0.66	0.49	hatev	aghs	-	0.60	0.62	0.59
	30	trac	oag	-	0.56	0.65	0.61	w&h	hs	0.30	0.53	0.61	0.55	offer	tox	-	0.60	0.61	0.60
	31	storm	hs	-	0.53	0.63	0.58	hatev	hs	0.35	0.52	0.60	0.56						
	32							storm	hs	0.30	0.50	0.60	0.55						

Table 5.3: Continuation of Table 5.2.

		1						2						3					
		train ->	**kaggl obsc**					**train ->**	**w&h sex**					**train ->**	**storm hs**				
		model ->		**svm**	**ft**	**be**	**al**	**model ->**		**svm**	**ft**	**be**	**al**	**model ->**		**svm**	**ft**	**be**	**al**
Testing	1	fount	abus	-	**0.88**	**0.90**	**0.87**	ami	sex	0.35	0.66	**0.73**	0.68	w&h	race	0.54	0.54	0.69	0.61
	2	fount	tox	-	**0.85**	**0.89**	**0.86**	david	offe	0.19	0.53	0.69	0.61	kaggl	insu	0.50	0.51	0.65	0.60
	3	offen	tox	-	**0.71**	**0.77**	**0.82**	david	tox	0.15	0.48	0.65	0.57	kaggl	hs	0.51	0.53	0.64	0.57
	4	david	tox	-	0.55	**0.73**	**0.75**	kaggl	obsc	0.49	0.56	0.64	0.56	kaggl	obsc	0.50	0.51	0.63	0.59
	5	david	offe	-	0.58	**0.73**	**0.74**	kaggl	insu	0.49	0.57	0.63	0.56	kaggl	tox	0.49	0.50	0.61	0.58
	6	ami	sex	-	0.67	**0.71**	0.69	kaggl	stox	0.50	0.59	0.60	0.53						
		train ->	**ami sex**					**train ->**	**w&h hs**					**train ->**	**hatev hs**				
		model ->		**svm**	**ft**	**be**	**al**	**model ->**		**svm**	**ft**	**be**	**al**	**model ->**		**svm**	**ft**	**be**	**al**
Testing	7	kaggl	insu	0.62	0.65	0.49	0.66	ami	sex	0.35	0.66	**0.70**	0.61	kaggl	insu	0.57	0.59	0.49	0.64
	8	kaggl	obsc	0.63	0.65	0.49	0.67	kaggl	insu	0.49	0.56	0.66	0.57	kaggl	obsc	0.57	0.60	0.49	0.65
	9	kaggl	tox	0.59	0.63	0.48	0.65	kaggl	obsc	0.49	0.56	0.66	0.57	kaggl	tox	0.54	0.61	0.48	0.66
	10	w&h	sex	0.56	0.56	0.44	0.62	david	offe	0.19	0.56	0.63	0.54	storm	hs	0.50	0.55	0.44	0.64
	11	hatev	aghs	0.61	0.59	0.44	0.57	david	tox	0.15	0.51	0.63	0.51	fount	abus	0.52	0.56	0.44	0.64
	12	fount	abus	0.54	0.60	0.44	0.62	kaggl	tox	0.48	0.55	0.62	0.55	fount	tox	0.51	0.57	0.42	0.64
	13	offen	tox	0.47	0.62	0.43	0.58	kaggl	hs	0.51	0.52	0.62	0.54	ami	sex	**0.73**	**0.99**	0.36	**0.84**
	14	fount	tox	0.52	0.59	0.42	0.61	kaggl	stox	0.51	0.53	0.61	0.54	davic	offe	0.68	0.62	0.19	**0.70**
	15	david	offe	**0.74**	0.65	0.19	**0.71**	storm	hs	0.47	0.57	0.61	0.57	davic	tox	0.63	0.58	0.15	0.67
	16	david	tox	0.69	0.61	0.15	0.66	hatev	hs	0.38	0.52	0.60	0.59						
		train ->	**fount hs**					**train ->**	**kaggl hs**					**train ->**	**david hs**				
		model ->		**svm**	**ft**	**be**	**al**	**model ->**		**svm**	**ft**	**be**	**al**	**model ->**		**svm**	**ft**	**be**	**al**
Testing	17	w&h	race	0.47	0.54	**0.77**	0.51	fount	hs	-	0.51	0.64	0.60	w&h	race	0.47	0.48	0.47	0.62
	18	storm	hs	0.44	0.47	**0.71**	0.48												
	19	w&h	hs	0.40	0.44	0.63	0.53												
	20	trac	oag	0.43	0.45	0.62	0.47												
	21	kaggl	hs	0.50	0.55	0.60	0.58												
		train ->	**hatev aghs**					**train ->**	**w&h race**										
		model ->		**svm**	**ft**	**be**	**al**	**model ->**		**svm**	**ft**	**be**	**al**						
T	22	kaggl	stox	0.53	0.61	0.50	0.60	storm	hs	0.47	0.46	0.66	0.55						

sures feature importance by observing how random re-shuffling of each predictor influences model performance, without changing the distribution of the variable. Before using this model, I normalize data with the Z-score method (Kreyszig, 1960).

5.1.6 Discussion

In this section, I discuss in detail the outcome of the experiments.

Outcome of Intra-dataset Classification

As can be observed in Figures 5.1 and 5.2, in general (and as expected from previous works), SVM is the model that performs worst, with some exceptions. Thus, in general, it performs poorer than fastText, except for the ‘abusive’ category of the *Founta* dataset, where both scored equally and the ‘offense’ category of the *Davidson* dataset, where SVM is slightly better. It also performs worse than ALBERT for all cases, except for ‘hate speech’ in the *Hateval* dataset, where it is slightly better.

FastText performs generally worse than BERT and ALBERT, which is in line with d’Sa et al. (2020) and Uglow et al. (2019), who also report a poorer performance of fastText compared to BERT.

Even though BERT and ALBERT achieve an overall better performance than the baseline models, BERT’s performance is not good (lower than 0.52) in some categories: ‘hate speech’ in *Davidson*, ‘sexism’ in *Ami*, both categories in *Hateval*, ‘covert aggression’ in *TRAC* and ‘hate speech’ in *Stormfront*. In these cases, both SVM and fastText, or at least one of them, obtain a better performance than BERT. This may be explained by the fact that BERT is unstable on smaller datasets Devlin et al. (2019). BERT is more unstable than ALBERT and fastText, both in terms of the same category (see, e.g., ‘hate speech’, ‘sexism’) Figure 5.1, and same dataset (*TRAC*, *Hateval*, *Davidson*) Figure 5.2. However, BERT also achieves the best performance on the largest number of categories.

As illustrated in Figure 5.1, from the categories present in more than one dataset, ‘toxicity’ proves to be the easiest category to classify, fol-

Table 5.4: Dataset and model features used for predicting model performance.

Feature	**Description**
Training dataset	One of *W&H*, *Davidson*, *Ami*, *TRAC*, *Hateval*, *Kaggle*, *Stormfront*, *Offenseval*, or *Founta*.
Training dataset size	Number of instances used for training.
Training dataset percentage of positive instances	Number of instances in the positive class divided by the number of instances used for training.
Original F1	Performance of the model when trained and tested with data from the same class, as provided in Figures 5.1 and 5.2.
Training category	All the distinct standardized categories in Table 4.1.
Training social network	Social network of the data: Twitter, Facebook, Wikipedia or Stormfront.
Test dataset	One of the same dataset list as in the training dataset.
Test dataset size	Number of instances in the test dataset.
Test category	One of the same category list as in the training category.
Test social network	One as in the same social network list as in the training dataset.
BERT, ALBERT, fastText or SVM	Model used to classify.
Is same category	Boolean indicating whether training and test sets belong to the same category.
Is same social network	Boolean indicating whether training and test set belong to the same social network.
Train and test set proportion	Training dataset size divided by the test dataset size.
Vocabulary-train	After removing stop words (by using NLTK) and keeping only distinct words, I compute the percentage of words that is present in the positive class of the test set.
Vocabulary-test	With the same procedure as for 'Vocabulary test present training', the percentage of distinct words from the positive class that are present in the positive class of the training set is computed.

lowed by 'sexism' and then 'hate speech', the latter one with more unstable results. From the dataset-specific categories, 'abusive' is the easiest to classify, followed by 'obscenity', 'offensive', 'racism', and 'aggression'. The remaining categories show worse performance.

According to Figure 5.2, among the datasets with more than one category, *W&H* shows good and stable results for all categories, while *TRAC* and *Hateval* show worse performances. For the other datasets, the performance varies from category to category. *Davidson* and *Founta* both show good performances, except for 'hate speech'. *Kaggle* shows good performance for 'obscene', 'toxicity' and 'insult', but worsens for 'hate speech', 'threat' and 'severe toxicity'. Among the datasets with only one category, *Ami*, *Offenseval*, and *Stormfront* have the best scores.

When comparing the results in Figures 5.1 and 5.2 with the figures reported in the literature (see Table 5.1), ALBERT and BERT models achieve similar performance as reported in other transformer-based works (such as, e.g., (Swamy et al., 2019)) for the classification of, e.g., 'hate speech' in *W&H*, 'toxicity' in *Davidson*, 'aggression' in *TRAC* and 'toxicity' in *Founta*. These models outperform works that do not use transformers; see, e.g., Pamungkas and Patti (2019) for classification of 'sexism' in *Ami*; de Gibert et al. (2018) for classification of 'hate speech' in *Stormfront*; Basile et al. (2019) for classification of 'hate speech' in *Hateval*, and van Aken et al. (2018) for classification of 'toxicity' in *Kaggle*.

Outcome of Inter-Dataset Classification

The results of the inter-dataset classification experiment provide some interesting insights with respect to both models and datasets. Both are discussed in the following subsections.

Discussion of the Models Table 5.2 shows that BERT and ALBERT models generalize better more often than fastText and SVM. Thus, the results for fastText are worse than for BERT and/or ALBERT, except for the generalization of *fount.tox* to *kaggl.obsc* (F1 of 0.82) and *kaggl.insu* (0.78), and *kaggl.insu* to *fount.abus* (0.70). For this last case, it is the only model capable to generalize. SVM performs generally worse than fastText, BERT and ALBERT, being better only when trained on *offen.tox* and tested on *kaggl.obsc* or *kaggl.insu* (0.81 and 0.74 respectively), or trained on *ami.sex* and tested on *david.offe* (0.74). In contrast to fastText, SVM does not show a distinct generalization potential, i.e., it is capable to generalize only if BERT, ALBERT or fastText also are.

BERT generalizes best for almost all the datasets; ALBERT is close to BERT many times and is even better for the models trained with *hateval.hs* and *ami.sex*.

In general, transformer-based models generalize better (this had already been shown by Swamy et al. (2019) while using BERT), while other models are less suitable for this task; see, e.g., Waseem et al. (2018); Gröndahl et al. (2018); Karan and Šnajder (2018). The carried experiments also show that the generalization capability of a model equally depends on the chosen dataset, and, even more importantly, on the targeted categories; see below.

Discussion of the Datasets and Categories BERT models that are trained on the category ‘toxicity’ of a dataset (*Offenseval*, *Davidson*, *Founta*, and *Kaggle*) generalize well over the same category of the other test sets; see, when trained on *offen.tox*: (0.91;0.83;0.76);[9] on *david.tox*: (0.91;0.81;0.85), on *fount.tox*: (0.84;0.87;0.77); and on *kaggl.tox*: (0.92;0.85;0.78). This shows that ‘toxicity’ is homogeneous across different datasets.

‘Offensive’ and ‘abusive’ are also consistently predicted well and when a model is trained on one of them, it also predicts well ‘toxicity’. E.g., when trained on *david.offen*, BERT predicts well *fount.tox* (0.89) and *kaggl.tox* (0.79), and ALBERT predicts well *offen.tox* (0.79). BERT also predicts well *david.offen* when trained on these datasets (0.73;0.72;0.73); and when trained on *fount.abus*, it predicts well *offen.tox* (0.83), *kaggl.tox* (0.85), and *david.tox* (with a borderline result of 0.69). *fount.abus* is also predicted well when BERT is trained on these datasets (0.88;0.90;0.89).

Categories such as ‘abusive’, ‘offensive’, ‘aggression’ and ‘toxicity’,

[9] If not mentioned otherwise, I cite the BERT figures.

whose definitions tend to overlap, generalize well between each other, which indicates that these labels are conceptually similar or that they represent the same phenomenon. Thus, when trained on *david.offen*, BERT predicts well *fount.abus* (0.89), *kaggl.tox* (0.79), and *offen.tox* (0.79). The prediction of *david.offen* is also accurate when BERT is trained on each of these datasets (0.71;0.73;0.72). The same holds for training on *fount.abus* and predicting *david.offen* (0.71) and the reverse (0.89), for training on *fount.abus* and testing on *kaggl.tox* (0.85) and the reverse (0.90), and for training with *fount.abus* and testing on *offen.tox* (0.83) and the reverse (0.88).

Datasets that include the categories ‘abusive’, ‘offensive’, ‘aggression’ or ‘toxicity’ also include ‘obscene’. ‘Obscene’ from *Kaggle* (*kaggl.obsc*) as training set obtains good results in different cases: ‘toxicity’ by BERT (see fount.tox: 0.89) and ALBERT (see *offen.tox*: 0.82, and *david.tox*: 0.75); ‘offensive’ by ALBERT (*david.offen*: 0.74); and ‘abusive’ by BERT (*fount.abus*: 0.90). Another category that is related to ‘abusive’, ‘offensive’, ‘aggression’ and ‘toxicity’ is ‘insult’. When using any of the former for training, models generalize reasonably well over *kaggl.insu* (see, for BERT *fount.tox*: 0.77, *david.offen*: 0.75, and *fount.abus*: 0.81, *david.tox*: 0.77, for SVM *offen.tox*: 0.74), proving that insults are also commonly subsumed by these categories. In view of the co-occurrence of ‘obscene’ and ‘insult’ with ‘toxicity’ reported in Fortuna et al. (2020), these generalizations are not surprising. Additionally, *Ami* ‘sexism’ seems to contain many insults: BERT generalizes well over *Ami.sex* when trained on *Ka* ‘insult’ (see *kaggl.insu*:0.74).

BERT trained on *TRAC* ‘overt aggression’ or on ‘aggression’ is capable of predicting ‘abusive’, ‘offensive’, and ‘toxicity’-related categories. Thus, training on *trac.oag*, predicts well *founta.abus* (0.80), *fount.tox* (0.82), *kaggl.tox* (0.79) and *offen.tox* (0.72). This is similar to when it is trained on *trac.aggr* (0.80; 0.84; 0.66 and 0.74, respectively), which is to be expected since both *TRAC* ‘aggressive’ and *TRAC* ‘overt aggressive’ datasets share data. BERT also generalizes better over *kaggl.insu* when trained on *trac.oag* (0.70). On the other side, when these predicted categories are used for training, the models generalize worse over *TRAC* 'overt aggressive', and they do not generalize over *TRAC* 'covert aggresive' or *TRAC* 'aggressive'. This can be due to the fact that *TRAC* contains covert and overt instances of harmful behavior in general. As a result, when models trained on *TRAC* are applied to other datasets, they can still detect and flag the positive instances of more explicit harm. However, models trained on other datasets struggle to deal with data that mostly contain covert aggression.

Table 5.2 also shows that models trained on the 'hate speech' category of the different datasets generalize much worse. Thus, BERT trained on *Founta* 'hate speech' generalizes over *Stormfront* with a worse performance, and poorly over *W&H* (trained on *fount.hs*, BERT's performance on *storm.hs* is 0.71 and on *W&H.hs* 0.63).

Poor performance is also observed for BERT trained on *Kaggle* 'hate speech' over *Founta* (trained on *kaggl.hs*; BERT's performance on *fount.hs* is 0.64. For the remaining datasets with hate speech categories (*W&H*, *Davidson*, *Hateval*, and *Stormfront*) the achieved generalization performance is even worse.

However, it is to be noted that in the case of more specific hate speech categories, a better generalization is observed; see, e.g., the generalization over *W&H* 'racism' when trained on *Founta* 'hate speech' and over *Ami* 'sexism' when trained on *W&H* 'sexism' (trained on *fount.hs*, BERT's performance on *W&H.race* is 0.77, and when trained on *W&H.sex*, its performance on *Ami.sex* is 0.73), which opposes Arango et al. (2019) conclusion that *W&H* is a dataset with a low generalization potential due to its composition of messages of a low number of authors. Just the contrary: certain categories of this dataset generalize when classifying sexism from *Ami*'s dataset.

Furthermore, SVM trained on *Ami* 'sexism' generalizes over *Davidson*'s 'offensive' test set (0.74), which indicates that *Davidson* may contain sexist offensive content. On the other side, as expected, *Hateval* 'hate speech' generalizes extremely well over *Ami* 'sexism' because both datasets share data (Fersini et al., 2018). *Hateval* targets hate speech against immigrants and women, and Ami targets only hate speech against women (i.e., misogyny). So this second dataset will miss the immigrant hate messages from *Hateval*. It also generalizes over ‘offensive’ in *Davidson* (trained on *hatev.hs*, ALBERT’s performance on *david.offe* is 0.70 and trained on *hatev.hs*, fastText’s performance on *ami.sex* is 0.99). This suggests that these three categories may be related and some sexist hate speech may be present in the ‘offensive’ samples of *Davidson*. This is surprising since the *Davidson* dataset is annotated with respect to both ‘offensive’ and ‘hate speech’, and both categories are mutually exclusive in this case. This means that there is probably sexist content annotated in the *Davidson* dataset as ‘offensive’, but not as ‘hate speech’.

Comparison with Previous Studies In this subsection, I compare the outcome of the cross-dataset experiments with those reported in previous works mentioned in Section 3.1.3. First, it is difficult to compare the experiments conducted in this chapter with those presented in Waseem et al. (2018), as in the first binary classification are applied to all the standardized dataset categories while Waseem et al. (2018) used multiclass classification.

On the other hand, it is easier to compare the achieved results with the results of the other generalization studies, which tag all abusive language-related messages as ‘positive’ and the remaining messages as ‘negative’. For instance, in Gröndahl et al. (2018), the reported macro F1 scores are below 0.49 for all of the setups with the*Davidson* and W&H datasets. This performance is lower than what I reported above for the experiments. Karan and Šnajder (2018) also binarize the labels of the *W&H*, *TRAC*, and *Kaggle* datasets. They report a generalization performance across the different categories of F1 scores < 0.48. Better scores are achieved when training BERT with *TRAC*’s ‘aggression’ and testing on *Kaggle*’s ‘toxicity’ (F1 score of 0.66) and *W&H*’s ‘hate speech’ (F1 score of 0.61); and when training BERT on *W&H*’s ‘hate speech’ and testing on *Kaggle*’s ‘toxicity’ (F1 score of 0.62).

Swamy et al. (2019) binarize the *W&H*, *Davidson*, *Offenseval*, and *Founta* datasets. Since they also use BERT, in the majority of the cases, their and the results in this chapter are comparable. In some cases, the second achieved a slightly higher performance. This is the case when BERT is trained on *Offenseval*'s 'toxicity' and tested on *Founta*'s 'toxicity'(0.91) and *Davidson*'s 'toxicity'(0.76); when it is trained on *Davidson*'s 'toxicity' and tested on *Founta*'s 'toxicity' (0.91) and when the training and test sets are reversed (0.77). The figures are better when BERT is trained on *Davidson*'s 'toxicity' and tested on *Offenseval*'s 'toxicity' (0.81); when it is trained on *Founta*'s 'toxicity' and tested on *Offenseval*'s 'toxicity' (0.84) and *W&H* 'hate speech' (0.67); and when it is trained on *W&H*'s 'hate speech' and tested on *Davidson*'s 'toxicity' (0.63). The overall (slightly) better performance in these experiments may be due to the fact that $\text{BERT}_{\text{LARGE}}$ is used, while Swamy et al. (2019) use $\text{BERT}_{\text{BASE}}$.

In another study (Pamungkas and Patti, 2019), the authors binarize the categories of *W&H*, *Hateval*, and *Offenseval* datasets. For the experiments, they use LSTM and SVM, which both render a poorer performance than the one achieved in my experiments. Compare, for instance, the case when BERT is trained on *Offenseval*'s 'toxicity' and tested on *Hateval*'s 'hate speech', and when it is trained on *W&H*'s 'hate speech' and tested on *Hateval*'s 'hate speech' (both with F1 = 0.60).

The above comparison of the outcome of the experiments with previous studies shows that the tested deep models perform, in general, better. As a look at Swamy et al. (2019) furthermore shows, different variants of BERT (in this case, $\text{BERT}_{\text{BASE}}$ vs. $\text{BERT}_{\text{LARGE}}$) also perform differently with larger models performing better.

Outcome of Model Performance Classification

The model performance classification aims to answer the open research question "Which model and dataset characteristics are important for generalization, after all?". Table 5.5 displays the feature importance of the 16 features obtained in the Random Forest classification experiment (F1 score of 0.64).[10]

Four features are most informative. The importance of 'original F1' (0.22) shows that for cross-dataset generalization it is relevant to start with a model that performs well in an intra-dataset scenario–something which

[10] See Table 5.4 for a descriotion of these features.

Table 5.5: Random Forest model feature importance.

features	**Imp.**
original F1	0.22
train category - toxicity	0.11
Vocabulary test	0.10
fastText	0.10
Train and test set proportion	0.07
Vocabulary train	0.06
test concept - toxicity	0.06
Training dataset size	0.06
BERT	0.05
test concept - hate speech	0.05
train concept - overt aggression	0.04
test concept - offense	0.03
Training dataset percentage of positive instances	0.03
SVM	0.03
Test dataset size	0.02
ALBERT	0.02
train concept - hate speech	0.02
test dataset - davidson, founta	0.02
test sn - twitter, facebook	0.01
Is same social network	0.01
train dataset - trac, founta	0.01
train concept - insult, obscene	0.01
test dataset - trac, ami	0.01
test concept - overt aggression, abusive, sexism, obscene	0.01
Remaining features	0.00

has been ignored in previous studies. ‘Vocabulary-test’ (0.10) proves also to be relevant. It is also worth pointing out that the generalization relies more on ‘Vocabulary-test’ (0.10) and less on ‘Vocabulary-train’ (0.06). It is advantageous to have in the training set a higher share of vocabulary of the test data in order to avoid “out-of-vocabulary” words. It seems to be of no advantage to have in the test set a high percentage of vocabulary appearing in the training set.

FastText (0.10) proves to be a good predictor of a poor generalization potential. BERT (0.05), SVM (0.03), and ALBERT (0.02) have lower relevance, which suggests that they add little to the already considered fastText variable.

Regarding categories, the feature ‘Train category - toxicity’ is also of relevance (0.11). This is in line with Karan and Šnajder (2018), who already pointed out that different dataset categories could affect generalization. In this case, ‘toxicity’ as a training set category led to good performance. One could expect that the feature ‘Test category - toxicity’ is also of relevance; however, it seems not to offer any further information to ‘Train category - toxicity’, since generalization profits from the use of ‘toxicity’ in both the training and test sets. The other category-related features contribute less, no matter whether they are used for training or testing. This also applies to all datasets and social network features.

With the works of Karan and Šnajder (2018) and Swamy et al. (2019) in mind, the dataset size-related features are expected to be of high importance. However, ‘Train and test set proportion’ obtained only 0.07, ‘Training dataset size’ 0.06, ‘Training dataset percentage of positive instance’ 0.03, and ‘Test dataset size’ 0.02. This might be due to the fact that dataset size-related variables have actually low variability in this and previous studies, but, rather, depend on the considered abusive language datasets. This would imply that both the carried experiments and previous research in the field are not the most suitable for assessing the effect of the dataset size-related variables. In this regard, the experiments in this chapter provide a further insight that the presence of categories with different performance in the same dataset makes it even more difficult to find possible correlations between dataset size-related variables and performance;

Table 5.6: Correlation between intra and cross model performance and dataset size features.

	Intra-Dataset	**Cross-Dataset**
	F1 macro	F1 macro
training data total size	0.14	0.01
training data total positive	0.31	0.30
training data percentage positive	0.07	0.16
testing data total size	0.15	0.24

see Table 5.6.

5.2 Perspective API Generalization Experiment

In order to analyse the generalization potential of a state-of-the-art model over the considered datasets and their categories, *Perspective API* is used.

5.2.1 Procedure

For each standardized category in the datasets,[11] I evaluate how well the Perspective API classifier is able to identify it and distinguish it from non-harmful messages. In other words, binary classification is performed using only the messages belonging to the analyzed category and the messages marked in the dataset as 'non-toxic', 'aggressive' or in any way 'abusive'. For the evaluation of the performance of the classifier on each dataset, the F1 metric is used.

5.2.2 Results and Discussion

Figure 5.3 shows the results of the classification experiment with Perspective API. The results reveal that the performance of the classifier has

[11] Due to the API quota limits, 20% of the *Kaggle* dataset are sampled in a total of 31.914 messages.

a huge variation depending on the category. The classifier is better at identifying ‘toxicity’, ‘offense’, followed by ‘obscene’, ‘insult’, ‘misogyny-sexism’, ‘hate speech’, and worse at identifying ‘aggression’, ‘racism’, ‘severe toxic’ and ‘threat’.

It is interesting to notice that despite the fact that the Perspective API classifier draws upon the same categories as used in the *Kaggle* dataset for annotation, the classifier seems to handle better ‘misoginy-sexism’ than ‘racism’.

Additionally, for the same category, the performance has high variability across datasets. For instance, for the hate speech category, the classifier shows a higher F1 for the Davidson dataset than for the Hateval dataset. The performance is even worse for the *Kaggle* dataset. This confirms that each dataset provides its own flavor of hate speech. For the general categories, ‘toxicity’ and ‘aggression’, the classifier achieves a higher F1 on the Davidson dataset than on the *Kaggle* dataset. The performance is even worse for the *TRAC* ‘aggression’. This means that, indeed, the ‘aggression’ category as used in the *TRAC* dataset cannot be compared and merged with the ‘toxicity’ category.

Also, the comparison with the performance on the *Kaggle* dataset shows that Perspective API performs better when applied to categories with more instances in the dataset such as ‘toxic’, ‘obscene’ and ‘insult’ and worse when applied to smaller categories such as ‘hate speech’, ‘severe toxic’ and ‘threat’. This indicates that the sampling procedure has a direct impact on the performance of the classifier, as better-represented classes are clearly better identified.

This experiment also confirms Kumar et al. (2018)’s observation that covert aggression (‘TRAC-CAG’) is recognized worse than overt aggression (‘TRAC-OAG’).

This experiment, which implied the use of the Perspective API classifier, shows that even when datasets use very generic categories, their diverging definitions, data samples or inconsistent annotation may lead to diverging classifier performance – as, e.g., in the case of ‘aggression’ from the TRAC dataset and ‘toxicity’ from the Kaggle dataset.

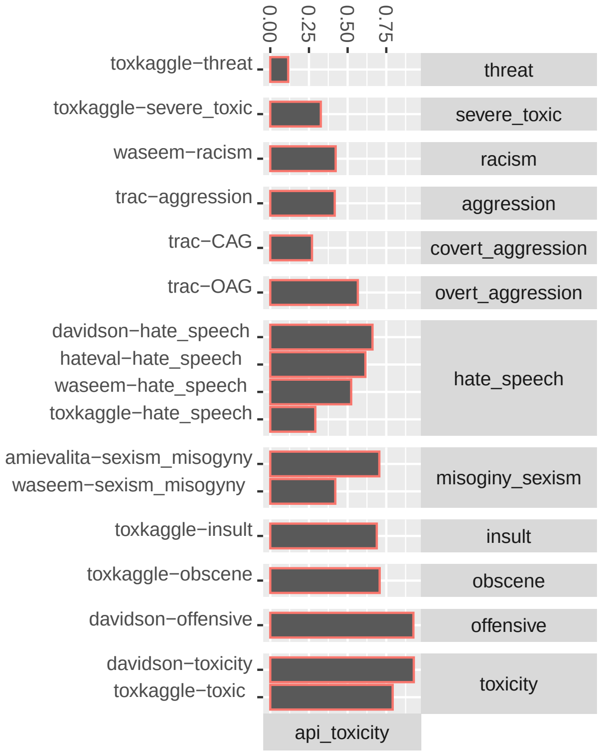


Figure 5.3: Toxicity’s Perspective API classification performance by category (F1 metric).

5.3 Chapter Discussion and Summary

The results of these experiments have some clear implications for the definition of the categories and model evaluation in the context of abusive language and hate speech in particular. In what follows, I summarize the main insights that these experiments provide.

Definitorial Implications

Hate speech is often conceptualized under umbrella terms such as “abusive language”, “offensive language”, or “toxicity” (e.g., Poletto et al., 2021; Jigsaw, 2019a). However, coarse-grained categories, like ‘toxicity’, should be used carefully. It is true that such categories lead to a good cross-dataset generalization in the experiment in Section 5.1.4 and with this in mind, it would be possible to conclude that broader coverage terms work well. However, the Perspective API experiment also provides evidence that the toxicity classifier has high variability, which depends on the targeted ‘toxicity’ subcategory: for instance, concepts such as ‘obscene’ are better detected than ‘hate speech’. If coarse-grained general categories may serve, it is necessary to clearly define and quantify the particular phenomenon that a general category in a dataset is supposed to cover. Subcategories should be annotated such that an error analysis on the model performance can be conducted and it can be assessed whether models equally detect all the subcategories. Given the results of the Perspective API experiment, it is important to prevent a concept drift where hate speech cedes prominence to more generic concepts such as ‘toxicity’. Otherwise, we risk to provide general classifiers that may have a general good performance but will fail for hate speech.

The results suggest that in the case of ‘hate speech’, more fine-grained categories would be more appropriate. When models are trained and tested on fine-grained categories such as ‘sexism’ or ‘racism’, better levels of generalization are achieved. Thus, the use of categories such as ‘hate speech’ does not help in terms of generalization. They are also likely to contain message samples that largely vary across the datasets with respect to content and style and thus do not serve well as training categories. This further buttresses the argumentation for a more fine-grained classification, e.g., in Fortuna et al. (2019); Salminen et al. (2018). A more fine-grained classification implies that during the dataset compilation and category definition phase, specific phenomena that define each category should be identified. Thus, if during the dataset compilation, only messages targeting sexism and racism are collected, a model trained on this dataset will not generalize well over a hate speech dataset that targets, for instance, homophobia. Fine grained categories of hate speech (e.g., 'sexism') also may have a different meaning in regard to different cultural contexts, and this is an aspect worth to consider in future experiments.

Implications for Model Evaluation

Historically and conventionally, in NLP, model generalization was restricted to a single data sample: generalization from training to test set. However, since both sets of data are expected to be homogeneous, i.e., collected simultaneously or with overlapping authors and topics for distinct messages, and given the rate at which new data is generated on social media, the utility of models that cannot generalize beyond that scope may also be questioned. In the case of the provided experiments, if only these techniques would have been considered, the community could conclude erroneously that hate speech classifiers are ready for a generalized deployment and usage.

With the process of evaluating models using a cross-dataset scenario, it is feasible to show how generalization can be designed from a broader perspective. What is of interest when deploying models is the degree to which models can categorize data from various sample sets. Or, to put it in another way, how do classifiers perform when a different sample of data is used for testing? Furthermore, what factors contribute to the greater generalizability of models? These are key issues, since classifiers in ML and AI are frequently not accompanied by the description of the environment in which they are intended to generalize and operate.

Cross-dataset testing helped to clarify some of these issues. By applying a random forest, it is possible to find out the factors contributing most to generalization. While models trained on global umbrella categories (‘toxicity’, ‘abusive’ or ‘offensive’ language) generalize more easily, general hate speech classifiers do not do well in a cross-dataset setup. Another factor that shows to be beneficial is to have a larger proportion of the test data vocabulary in the training set in order to prevent “out-of-vocabulary” words. This shows that categorization models continue to depend on the employed vocabulary. Furthermore, geographical, cultural and socio-political values and factors underneath each definition of hate speech may also justify why different hate speech classifications are incompatible and do not cross-generalize.

In the next chapter, I build upon the results presented so far in this thesis and continue to elaborate on the issues of applying NLP classification to the problem of hate speech with an eye on the future.

Chapter 6

PRESENT AND FUTURE OF HATE SPEECH CLASSIFICATION

Given the extent of the problems identified in Chapters 3 to 5, it is timely to assess whether the classification conventions and practices that have been followed so far are indeed appropriate for the task of hate speech classification. The objective of this chapter is to explore from a theoretical point of view the future of the application of supervised classifiers to hate speech detection (see Goal 7) and determine whether the task should be continued or discarded. The majority of the content of this chapter is derived from the second section of Fortuna et al. (2022).

6.1 Continuing to Combat Online Hate Speech Through the Use of Classifiers

A prerequisite for the application of supervised ML to hate speech detection is the solution of the problems identified in the previous chapters. Overcoming the identified challenges will require shifting research practices and prove improvements at the level of definitions, annotation and model evaluation.

6.1.1 Improving Definitions

Accounting for Non-neutrality of Definitions. While several interpretations for hate speech reflect a variety of societal values, the NLP community assumed a unified concept of hate speech and neutrality when defining it. A solution that would allow to consider more than one definition is *model framing*: narrow down definitions and clearly define the set of values and goals underneath each classifier. Deploying a supervised classifier would require specifications referring to a legal framework, the social media platform(s) of application, the concrete minorities to protect for a specific language, and a specific and well defined generalization context for the model. Such *model framing* can help address the issues surrounding universality and can provide space for researchers to consider how their choices have political implications for what speech is sanctioned.

Text Contextualization. Supervised ML models for hate speech primarily operate on texts. However, whether a text is to be qualified as hate speech is highly context dependent (Waseem et al., 2018). For instance, whether a word is used as a slur or as a reclaimed term depends on the identity of the speaker and the phrasal and social contexts in which it is uttered. The primary means of approximating *conversational context* has been in prior work through the use of conversation threads and user metadata (e.g. Mosca et al., 2021; Gao and Huang, 2017). Adding such data accounts only for a part of the context and excludes the societal background present during the interaction between individuals. For example, annotators hold prior knowledge on the histories, social hierarchies, conflicts, or stereotypes concerning the groups addressed in the annotation of a document. This knowledge is vital when manually analyzing the text. Hate speech detection research would therefore benefit from developing methods that allow for social knowledge to be encoded in the model prior to the training phase.

6.1.2 Improving Annotation

Having Representative Sampling Procedures. Given that sampling methods are used to ensure rich hate speech datasets, models are often trained on data distributions that significantly vary from real-world occurrences of hate speech. This has the disadvantage that datasets are biased and the models do not generalize when applied to new data. To address this concern, future data collection efforts should seek to reflect the real distribution of hate speech in the media under analysis. Obviously, this also implies that the models must be improved to be able to cope with highly unbalanced data.

Having Representative Annotations. The common NLP practice of having a small number of annotations for each document, which are used to compute IAA and assign labels based on a majority vote depends on the annotators at work because the interpretation of what constitutes hate is in general a highly subjective question. While subjectivity is inherent, I propose that researchers do not use binary annotation with a reduced number of annotators. One possible alternative would be to investigate procedures more similar to applying scales (DeVellis and Thorpe, 2021) and measure different dimensions of hate speech. Scales have previously been used in the social sciences for asking subjective questions, and could provide new possibilities for hate speech research. When applying scales, instead of having one question (e.g., “Does this text contain hate speech?”), it would be necessary to find the different dimensions of the construct (in this case “hate speech”) the scale seeks to measure. The different dimensions of the scale would need to undergo a validation procedure with a representative sample of the population.

6.1.3 Improving Model Understanding and Evaluation

Improving Model Generalization. I argue that the IID assumption is an impediment of the ability of models to generalize. The IID assumption is unlikely to hold since the process of creating datasets relies heavily on sampling strategies which over-emphasize data which may contain hate speech (see also the statement on the sampling procedures above). Additionally, the speed at which topics change in online social media (Hogan and Quan-Haase, 2010) makes it unclear whether time-bound data collections from social media can be IID to previous samples and the extent to which a model still applies to a certain data collection. Methods for determining whether or not various data samples are IID and whether or not models still apply must be devised.

Evaluating Hate Speech Models. In the last few years, research on hate speech detection has shown an increase in performance across a number of metrics. However, these increases do not reveal a full picture of the performance of a model. In fact, contemporary models display a superficial understanding of hate speech (see Section 3.1.3). It is therefore necessary for research on hate speech to consider new evaluation paradigms and metrics. There are multiple possible solutions to this problem: new metrics must center the abilities of the models to capture hate beyond the identification of frequently occurring tokens. An option is the creation of test suites that target potential areas of concern for models for detecting hate (e.g. Röttger et al., 2021) or to directly leverage training data for creation of hard-to-pass tests for ML models. Another avenue for improvement is to apply the strategy outlined in Chapter 5 and with cross-dataset experiments improve the model evaluation picture on generalization.

Improving Language Understanding. As mentioned above, models for hate speech seem to have superficial apprehension of language, which is incompatible with the complexity of the task (see Section 3.1.3). If models are over-fitting on spurious correlations and are incapable of language understanding (Arango et al., 2019; Bender and Koller, 2020), it begs the question whether we can rely on state-of-the-art classifiers for hate speech identification.

6.1.4 Additional Implications of Classification

Handling Classification Errors. For many NLP tasks, classification errors do not imply an immediate harm. In contrast, for hate speech detection, classification errors can result in significant immediate harm to people. False negatives can result in hateful speech being passed as acceptable, which can allow harmful content to remain unsanctioned (Oliva et al., 2020), while false positives can result in inoffensive speech being sanctioned. In light of these concerns, it is prudent for the NLP community and legislators to reflect on the ramification of classification errors. For instance, both NLP researchers and legislators should reflect on which entity is to be held responsible for such errors, and how victims of automated classification errors should be compensated.

Risking Marginalization. The implication of the specific ways in which hate speech classifiers under-perform has impacts on traditionally excluded groups. For instance, a non-identified sexist message is going to offend women while the topic of a conversation defending women's rights may be enough to activate a false positive. Several studies have shown how it is possible to silence minorities such as drag queens or African American English speakers via the use of NLP classifiers (e.g. Davidson et al., 2019; Oliva et al., 2020). Specifically, classifiers that are not able to accurately adjudicate content directed toward those groups risk increasing the costs for minorities to participate in online spaces.

Missing Deeper Explanations. Applying classifiers to hate speech detection is not informative regarding the relations between the groups involved in the aggression, and cognitive processes for perpetrators and victims. It is a methodology that helps little in understanding who are the targets of hate speech and which beliefs haters have about the targets.

Documenting the limitations discussed in this section and discussing ways to overcome them is needed for continuing to pursue the automatic classification of hate speech. However, even if all the identified difficulties are resolved, hate speech classifiers are instruments ready to be used by dominant groups to reinforce their power and are unlikely to help in the protection of minorities, as I will argue in the next section.

6.2 Hate Speech Classifiers as Tools Serving Colonialism

In this section, I explain why, from a systemic approach to hate speech (Gelber, 2021), researchers should abandon classification of hate speech since it supports colonialist relationships in the current days and reproduces power dynamics where minorities are kept in their status.

One interpretation for the term colonialism is the process of taking complete or partial governmental authority over another nation, filling it with settlers, and exploiting it, frequently by creating colonies (lin, 2021b). However, different definitions exist. According to the tradition of critical theory, the term "colonization" connotes a relationship of structural dominance aimed at preserving resources for dominant objectives (Mohanty, 1988), and the processes of colonization still take place in the present (Gregory, 2004). Recent research has been discussing how AI and ML are instruments capable of enabling and sustaining colonialism and are likewise reliant and reliable tools on colonial dynamics (Couldry and Mejias, 2021; Ricaurte, 2019). For instance, terms such as "digital colonialism" (Ricaurte, 2019) appears to express colonialism, power dynamics and oppression happening in the digital world.

Applying classifiers to detecting hate speech also plays a role in such digital colonial dynamics. This happens since developing hate speech classifiers implies control over discourses and access to privilege in different stages: the task of hate speech definition and annotation implies enumerating which groups are minorities and which utterances are unacceptable; after the training phase, the definition for hate speech is encoded in the model and future classifications are delegated to this entity which is not clear how to question; and furthermore, while AI and ML research communities create, design, define, comprehend, or utilize those tools, affected communities have no resources to understand their operation or paradigms, and no voice on their usage.

When analyzing hate speech detection models, a decolonial lens is useful. Using a decolonial lens means to acknowledge and to confront the ongoing impacts of colonialism and to analyze how, through the application of these algorithms, violence can operate (Ricaurte, 2019). Studies have shown how the application of hate speech classifiers can result in minorities being more policed online and their contributions more frequently identified as inappropriate and removed (Haimson et al., 2021). Also, classification errors often disproportionately affect marginalized communities. Thus, white supremacist content remains unsanctioned, while content from marginalized communities is removed (Davidson et al., 2019; Oliva et al., 2020). Regarding other errors models commit, in the second set of examples, Table 3.1 (examples 2.1-2.3), it is possible to examine how a state-of-the-art hate speech detection model responds to conversations around power dynamics. The explicit mentioning of gender and race prompts incorrect predictions by the model. Should this model be deployed, it would actively limit conversations around race, gender, and power dynamics more broadly. Such conversations are frequently held by communities that are marginalized, in an effort to identify, discuss, and seek to remedy their own marginalization. That is, the model is censoring conversations that are necessary to have, in order for the society to progress beyond contemporary forms of marginalization.

A systemic approach to hate speech (see Chapter 2) looks more promising in assuring less risks for minorities. This approach does not rely on the detection of hate, neither does it rely on the identification of specific vocabulary. It suggests a non-punitive, discourse-based strategy for responding to hate speech. In this thesis, I support the exploration of this direction. I defend that, to protect minorities, it is necessary to drop hate speech classification as a task, and the solution for AI and ML in the context of hate speech is to find other related tasks helping to build anti-hate narratives.

6.3 Chapter Summary

This chapter it is divided into two parts. In the first part, I elaborate on the steps that are necessary if classification is chosen as a means for solving the problem of online hate speech. These steps include accounting for the context-dependent plurality of the concept of hate speech, improving representativity of annotations, eliminating biases in the construction of datasets, and enhancing the capacity of models to represent language and new metrics that reflect this information. However, even if I propose new directions for the application of classifiers to the detection of hate speech, I do not expect that implementing any individual solution in isolation will result in ready-to-use classifiers. Rather, I emphasize the need for research to continuously reassess the risks that arise from methodological innovations for hate speech detection.

In the second part, I discuss the limitations of current hate speech classification models and how these limitations endanger minorities. Applying NLP classifiers has risks, and the present limitations support that these models perpetuate colonialist dynamics. Continuing to apply classifiers is incompatible with the anti-colonial and systemic approaches to fighting hate speech, which in my point of view are necessary to protect minorities and to let minorities write their own history.

In the next chapter, I conclude this thesis with a summary of the main research questions and findings as well as possibilities for future research.

Chapter 7

CONCLUSIONS AND FUTURE WORK

In this thesis, I have argued that current NLP practices for hate speech detection are unlikely to address the core concerns of hate speech detection, i.e., identify hate with minimal errors and protect marginalized communities. In the course of this thesis, I therefore encourage the NLP community to rethink their methodologies such that future developments reduce the risk for marginalized communities.

In Chapter 3, I presented how hate speech classifiers have been used following well-established conventions concerning task definition, data annotation and model training, testing and evaluation, but with limited results and facing profound challenges. The empirical experiments in Chapters 4 and 5 continued to elaborate on the limitations of current practices, conventions and models.

In Chapters 4, I further discussed hate speech-related terminology, concluding that improving the dataset quality in the area of hate speech detection is necessary and starting with definitorial challenges is a priority. Concepts are not thoroughly defined; the existing definitions ignore not only the subjectivity of this task, but also the geographical, cultural and socio-political context of hate speech. In general terms, previous investigations work with a single definition of hate speech assuming universality and disregarding the plurality of this concept.

In Chapter 5, I have shown that while models of global umbrella concepts ('toxicity', 'abusive' or 'offensive' language) or fine-grained categories ('sexism' or 'racism') generalize more easily, hate speech classifiers do not do well in a cross-dataset situation. Overall, I demonstrated that hate speech classifiers are not ready to generalize to new data. What has been shown to have better performance is to have classifiers for general concepts such as toxicity and having a larger proportion of the test data vocabulary in the training set in order to prevent 'out-of-vocabulary' words. This showed that categorization models continue to depend on the employed vocabulary. However, defining and categorizing hate speech is more complex than this. Judging a discourse as hate speech needs context and an examination of the power relations between certain groups and an awareness of how specific minorities are marginalized within the society.

Finally in Chapter 6, I have argued that current NLP practices for hate speech detection are unlikely to address the core concerns of hate speech detection, i.e., to identify hate with minimal errors and protect marginalized communities. I therefore called the NLP community to rethink their methodology such that future developments reduce risk for marginalized communities. I used a decolonial lens to highlight how much existing NLP methods fall short of safeguarding minorities and to guarantee that current technologies provide more advantages than harms.

Some of the ideas argued in this manuscript have been presented in the publications I co-authored during the period of the thesis (all but one as the first author):

- Fortuna, P., Dominguez, M., Wanner, L., and Talat, Z. (2022). Directions for NLP practices applied to online hate speech detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.*

- Fortuna, P., Soler-Company, J. & Wanner, L. (2022). Dataset annotation in abusive language detection. In *C. Strippel, S. Paasch-Colberg, M. Emmer & J. Trebbe (Eds.), Challenges and perspec-*

tives of hate speech analysis (pp. 369-391). Digital Communication Research.

- Fortuna, P., Soler-Company, J., & Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets?. *Information Processing & Management, 58(3), 102524.*

- Fortuna, P., Soler-Company, J., & Wanner, L. (2020, May). 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* (pp. 6786-6794).

- Shvets, A., Fortuna, P., Soler-Company, J., & Wanner, L. (2021, August). Targets and aspects in social media hate speech. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)* (pp. 179-190).

- Fortuna, P., Cortez, V., Ramalho, M. S., & Pérez-Mayos, L. (2021, August). MIN_PT: An European Portuguese Lexicon for Minorities Related Terms. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)* (pp. 76-80).

- Fortuna, P., Pérez-Mayos, L., Abura’ed, A., Soler-Company, J., & Wanner, L. (2021, August). Cartography of Natural Language Processing for Social Good (NLP4SG): Searching for Definitions, Statistics and White Spots. In *Proceedings of the 1st Workshop on NLP for Positive Impact* (pp. 19-26).

- Fortuna, P., da Silva, J. R., Soler-Company, J., Wanner, L., & Nunes, S. (2019, August). A hierarchically-labeled portuguese hate speech dataset. In Proceedings of the third workshop on abusive language online (pp. 94-104).

- Fortuna, P., Cruz, L.B., Maia, R., Cortez, V. & Nunes, S., 2021. Toxicity-Associated News Classification: The Impact of Metadata and Content Features. *ICWSM Workshops 2021*

- Fortuna, P., Soler-Company, J., & Nunes, S. (2019, June). Stop PropagHate at SemEval-2019 Tasks 5 and 6: Are abusive language classification results reproducible?. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 745-752).

- Fortuna, P., Ferreira, J., Pires, L., Routar, G., & Nunes, S. (2018, August). Merging datasets for aggressive text identification. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying* (TRAC-2018) (pp. 128-139).

- Fortuna, P., Bonavita, I., & Nunes, S. (2018, January). Merging datasets for hate speech classification in Italian. In *EVALITA@ CLiC-it*.

Future work on the problem of hate speech should discuss how AI, ML and NLP may still play a part in solving online hate speech by aiding with new tasks and data analysis. One possibility is to use NLP models to analyze and assist in the comprehension of content that has previously been identified as targeting minorities. One potential question to address is what are the targets' primary characteristics and aspects as stereotyped by the perpetrators (e.g. Shvets et al., 2021; Fraser et al., 2022). In this way, it would be possible to identify some stereotypes used against minorities and utilize them as a starting point for developing anti-hate interventions.

This thesis has limitations. While the analysis that I conducted allows for a deeper understanding of the problems with contemporary methods for hate speech detection, it focus on research rather than application. It therefore does not discuss how classification models are used in real-world content monitoring applications. This is left to future work.

However, I would like to emphasize the importance of this work to influence current practices on hate speech detection and analysis. By taking steps to document and address the limitations of contemporary methods for hate speech, this work provides evidence for the need to identify new avenues for improving the analysis of the online hate speech. Moreover, research needs to take steps toward ensuring that content moderation related technologies provide safer online spaces for marginalized communities by mitigating the prevalence of online hate speech.

Bibliography

(2021a). Cambridge dictionary. Available in https://dictionary.cambridge.org/, accessed last time in September 2021.

(2021b). Oxford dictionary. Available in https://www.lexico.com/, accessed last time in September 2021.

(2022). Uncovered: Online hate speech in the covid era. Available in https://www.ditchthelabel.org/research-papers/hate-speech-report-2021/, accessed last time in June 2022.

Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.

Agarwal, S. and Sureka, A. (2015). Using knn and svm based one-class classifier for detecting online radicalization on twitter. In *International Conference on Distributed Computing and Internet Technology*, pages 431–442. Springer.

Al Ramiah, A., Hewstone, M., Dovidio, J. F., and Penner, L. A. (2010). The social psychology of discrimination: Theory, measurement and consequences. *Making Equality Count: Irish and International Research Measuring Equality and Discrimination. Dublin, Ireland, The Equality Authority*.

Al Shalabi, L., Shaaban, Z., and Kasasbeh, B. (2006). Data mining: A preprocessing engine. *Journal of Computer Science*, 2(9):735–739.

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347.

Anzovino, M., Fersini, E., and Rosso, P. (2018). Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer.

Arango, A., Pérez, J., and Poblete, B. (2019). Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 45–54, New York, NY, USA. Association for Computing Machinery.

Arango, A., Pérez, J., and Poblete, B. (2020). Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, page 101584.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4(none):40 – 79.

Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., and Sanguinetti, M. (2019). SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Baumeister, R. F. and Finkel, E. J. (2019). Advanced social psychology: The state of the science.

Bender, E. M. and Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Biau, G. and Scornet, E. (2016). A random forest guided tour. *Test*, 25(2):197–227.

Blodgett, S. L., Barocas, S., III, H. D., and Wallach, H. M. (2020). Language (technology) is power: A critical survey of ”bias” in NLP. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. R., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5454–5476. Association for Computational Linguistics.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Brewer, M. B. et al. (1999). The psychology of prejudice: Ingroup love or outgroup hate? *Journal of social issues*, 55:429–444.

Bühlmann, P. (2012). Bagging, boosting and ensemble methods. In *Handbook of computational statistics*, pages 985–1022. Springer.

Chandrasekharan, E., Samory, M., Srinivasan, A., and Gilbert, E. (2017). The bag of communities: Identifying abusive behavior online with preexisting internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI’17, pages 3175–3187, New York, NY, USA. Association for Computing Machinery.

Chen, Y., Zhou, Y., Zhu, S., and Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80.

Chollet, F. and Allaire, J. (2018). Deep learning with r, manning publications.

Couldry, N. and Mejias, U. A. (2021). The decolonial turn in data and technology research: what is at stake and where is it heading? *Information, Communication & Society*, pages 1–17.

Davidson, T., Bhattacharya, D., and Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Davidson, T., Warmsley, D., Macy, M. W., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pages 512–515. AAAI Press.

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

DeVellis, R. F. and Thorpe, C. T. (2021). *Scale development: Theory and applications*. Sage publications.

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

Dovidio, J., Hewstone, M., Glick, P., and Esses, V. (2010). Prejudice, stereotyping, and discrimination: Theoretical and empirical overview. *The SAGE Handbook of Prejudice, Stereotyping and Discrimination*, pages 3–28.

d’Sa, A. G., Illina, I., and Fohr, D. (2020). BERT and fastText embeddings for automatic detection of toxic speech. In *SIIE 2020-Information Systems and Economic Intelligence*.

Fersini, E., Nozza, D., and Rosso, P. (2018). Overview of the evalita 2018 task on automatic misogyny identification (AMI). In Caselli, T., Novielli, N., Patti, V., and Rosso, P., editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Fiske, S. T. (2012). Journey to the edges: Social structures and neural maps of inter-group processes. *British Journal of Social Psychology*, 51(1):1–12.

Fortuna, P. Soler-Company, J. and Nunes, S. (2019). Stop propaghate at semeval-2019 tasks 5 and 6: Are abusive language classification results reproducible? In *Proceedings of the 13th International Workshop on Semantic Evaluation*.

Fortuna, P., Dominguez, M., Wanner, L., and Talat, Z. (2022). Directions for nlp practices applied to online hate speech detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computer Surveys*, 51(4):85:1–85:30.

Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., and Nunes, S. (2019). A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy. Association for Computational Linguistics.

Fortuna, P., Soler, J., and Wanner, L. (2020). 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 6788–6796, Marseille, France. European Language Resources Association.

Fortuna, P., Soler Company, J., and Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Inf. Process. Manag.*, 58(3):102524.

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., and Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 491–500. AAAI Press.

Fraser, K. C., Kiritchenko, S., and Nejadgholi, I. (2022). Computational modeling of stereotype content in text. *Frontiers in artificial intelligence*, 5.

Gao, L. and Huang, R. (2017). Detecting online hate speech using context aware models. In Mitkov, R. and Angelova, G., editors, *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017*, pages 260–266. INCOMA Ltd.

Gelber, K. (2021). Differentiating hate speech: a systemic discrimination approach. *Critical Review of International Social and Political Philosophy*, 24(4):393–414.

Glick, P. and Fiske, S. T. (2001). An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. *American psychologist*, 56(2):109.

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis lectures on human language technologies*, 10(1):1–309.

Gregory, D. (2004). *The Colonial Present: Afghanistan, Palestine, Iraq*. Blackwell Pub. Malden, MA.

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., and Asokan, N. (2018). All you need is "love": Evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, AISec'18, pages 2–12, New York, NY, USA. Association for Computing Machinery.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377.

Guermazi, R., Hammami, M., and Hamadou, A. B. (2007). Using a semi-automatic keyword dictionary for improving violent web site filtering. In Yétongnon, K., Chbeir, R., and Dipanda, A., editors, *Third International IEEE Conference on Signal-Image Technologies and Internet-Based System, SITIS 2007, Shanghai, China, December 16-18, 2007*, pages 337–344. IEEE Computer Society.

Haimson, O. L., Delmonaco, D., Nie, P., and Wegner, A. (2021). Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proc. ACM Hum. Comput. Interact.*, 5(CSCW2):466:1–466:35.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*. Springer Series in Statistics. Springer.

Hogan, B. and Quan-Haase, A. (2010). Persistence and change in social media. *Bulletin of Science, Technology & Society*, 30(5):309–315.

Hovy, E. H. and Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Lang. Linguistics Compass*, 15(8).

Jigsaw (2019a). Perspective api. Available in https://github.com/conversationai/perspectiveapi, accessed last time in November 2019.

Jigsaw (2019b). Toxic comment classification challenge: Identify and classify toxic online comments. Available in https://www.kaggle.com/c/jigsaw
-toxic-comment-classification-challenge, accessed last time in November 2019.

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J., Nithya, M., Kannan, S., and Gurusamy, V. (2014). Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks*, 5(1):7–16.

Karan, M. and Šnajder, J. (2018). Cross-domain detection of abusive language online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137, Brussels, Belgium. Association for Computational Linguistics.

Kawakami, K., Amodio, D. M., and Hugenberg, K. (2017). Intergroup perception and cognition: An integrative framework for understanding the causes and consequences of social categorization. In *Advances in experimental social psychology*, volume 55, pages 1–80. Elsevier.

Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., and Taboada, M. (2020). The SFU opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, 4(2):155–190.

Kreyszig, E. (1960). *Advances engineering mathematics*. Wiley Eastern.

Kühl, N., Goutier, M., Hirt, R., and Satzger, G. (2020). Machine learning in artificial intelligence: Towards a common understanding. *arXiv preprint arXiv:2004.04686*.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2018). 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. Association for Computational Linguistics.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

McIntosh, P. (1988). White privilege and male privilege: a personal account of coming to see correspondences through work in women's studies (1988). *On Privilege, Fraudulence, and Teaching As Learning*, pages 17–28.

McNamee, L. G., Peterson, B. L., and Peña, J. (2010). A call to educate, participate, invoke and indict: Understanding the communication of online hate groups. *Communication Monographs*, 77(2):257–280.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In Bengio, Y. and LeCun, Y., editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mohanty, C. (1988). Under western eyes: Feminist scholarship and colonial discourses. *Feminist Review*, 30(1):61–88.

Mosca, E., Wich, M., and Groh, G. (2021). Understanding and interpreting the impact of user context in hate speech detection. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 91–102, Online. Association for Computational Linguistics.

Mozafari, M., Farahbakhsh, R., and Crespi, N. (2019). A bert-based transfer learning approach for hate speech detection in online social media. In Cherifi, H., Gaito, S., Mendes, J. F., Moro, E., and Rocha, L. M., editors, *Complex Networks and Their Applications VIII - Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019, Lisbon, Portugal, December 10-12, 2019*, volume 881 of *Studies in Computational Intelligence*, pages 928–940. Springer.

Nobata, C., Tetreault, J. R., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content. In Bourdeau, J., Hendler, J., Nkambou, R., Horrocks, I., and Zhao, B. Y., editors, *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 145–153. ACM.

Noble, S. U. (2018). *Algorithms of oppression*. New York University Press.

Nozza, D. (2021). Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

Oliva, T. D., Antonialli, D. M., and Gomes, A. (2020). Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & Culture*, pages 1–33.

Olteanu, A., Castillo, C., Boy, J., and Varshney, K. R. (2018). The effect of extremist violence on hateful speech online. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 221–230. AAAI Press.

Pamungkas, E. W. and Patti, V. (2019). 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. Association for Computational Linguistics.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., and Hanna, A. (2020). Data and its (dis) contents: A survey of dataset development and use in machine learning research. *arXiv preprint arXiv:2012.05345*.

Pavlopoulos, J., Androutsopoulos, I., Thain, N., and Dixon, L. (2019). Convai at semeval-2019 task 6: Offensive language identification and categorization with perspective and bert. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 571–576.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2020). Scikit-learn support vector classification. Available at `https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html`, accessed last time October.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and

Édouard Duchesnay (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Pitsilis, G. K., Ramampiaro, H., and Langseth, H. (2018). Detecting offensive language in tweets using deep learning. *CoRR*, abs/1801.04433.

Poletto, F., Basile, V., Bosco, C., Patti, V., and Stranisci, M. (2019). Annotating hate speech: Three schemes at comparison. In Bernardi, R., Navigli, R., and Semeraro, G., editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, volume 2481 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Lang. Resour. Evaluation*, 55(2):477–523.

Prentice, S., Taylor, P. J., Rayson, P., Hoskins, A., and O'Loughlin, B. (2011). Analyzing the semantic content and persuasive composition of extremist media: A case study of texts produced during the gaza conflict. *Inf. Syst. Frontiers*, 13(1):61–73.

Pustejovsky, J. and Stubbs, A. (2012). *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. " O'Reilly Media, Inc.".

Rademacher, L. A. (2007). Approximating the centroid is hard. In *Proceedings of the twenty-third annual symposium on Computational geometry*, pages 302–305.

Rahman, M. M., Balakrishnan, D., Murthy, D., Kutlu, M., and Lease, M. (2021). An information retrieval approach to building datasets for hate speech detection. *arXiv preprint arXiv:2106.09775*.

Ricaurte, P. (2019). Data epistemologies, the coloniality of power, and resistance. *Television & New Media*, 20(4):350–365.

Roberts, S. T. (2019). *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.

Röttger, P., Vidgen, B., Nguyen, D., Talat, Z., Margetts, H. Z., and Pierrehumbert, J. B. (2021). Hatecheck: Functional tests for hate speech detection models. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 41–58. Association for Computational Linguistics.

Salminen, J., Almerekhi, H., Milenkovic, M., Jung, S., An, J., Kwak, H., and Jansen, B. J. (2018). Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 330–339. AAAI Press.

Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S.-g., Almerekhi, H., and Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10(1):1.

Sanguinetti, M., Comandini, G., Di Nuovo, E., Frenda, S., Stranisci, M., Bosco, C., Caselli, T., Patti, V., Russo, I., and Pisa, I. (2020). Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task. In *EVALITA*.

Santucci, V., Spina, S., Milani, A., Biondi, G., and Di Bari, G. (2018). Detecting hate speech for italian language in social media. In *EVALITA 2018, co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2263.

Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th An-*

nual Meeting of the Association for Computational Linguistics, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., and Smith, N. A. (2021). Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *CoRR*, abs/2111.07997.

Schmidt, A. and Wiegand, M. (2017). 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. Association for Computational Linguistics.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47.

Shvets, A., Fortuna, P., Soler, J., and Wanner, L. (2021). Targets and aspects in social media hate speech. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 179–190, Online. Association for Computational Linguistics.

SpaCy (2022). Word vectors and semantic similarity. Available in `https://spacy.io/usage/linguistic-features#vectors-similarity`, accessed last time in April 2022.

Swamy, S. D., Jamatia, A., and Gambäck, B. (2019). 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. Association for Computational Linguistics.

Tajfel, H., Billig, M. G., Bundy, R. P., and Flament, C. (1971). Social categorization and intergroup behaviour. *European journal of social psychology*, 1(2):149–178.

Tarasova, N. (2016). Classification of hate tweets and their reasons using svm.

Thylstrup, N. and Talat, Z. (2020). Detecting 'dirt' and 'toxicity': Rethinking content moderation as pollution behaviour. *Available at SSRN 3709719*.

Uglow, H., Zlocha, M., and Zmyslony, S. (2019). An exploration of state-of-the-art methods for offensive language detection. *CoRR*, abs/1903.07445.

van Aken, B., Risch, J., Krestel, R., and Löser, A. (2018). Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, Brussels, Belgium. Association for Computational Linguistics.

Van Dijk, T. A. (2013). Discourse, power and access. In *Texts and practices*, pages 93–113. Routledge.

Vidgen, B. and Derczynski, L. (2021). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):1–32.

Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., and Margetts, H. (2019). Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Vidgen, B., Thrush, T., Waseem, Z., and Kiela, D. (2021). Learning from the worst: Dynamically generated datasets to improve online hate detection. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1667–1682. Association for Computational Linguistics.

Vigna, F. D., Cimino, A., Dell'Orletta, F., Petrocchi, M., and Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook. In

Armando, A., Baldoni, R., and Focardi, R., editors, *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy, January 17-20, 2017*, volume 1816 of *CEUR Workshop Proceedings*, pages 86–95. CEUR-WS.org.

Waseem, Z. (2016). Are you a racist or am I seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Waseem, Z., Lulz, S., Bingel, J., and Augenstein, I. (2021). Disembodied machine learning: On the illusion of objectivity in nlp. *arXiv preprint arXiv:2101.11974*.

Waseem, Z., Thorne, J., and Bingel, J. (2018). Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In *Online Harassment*, pages 29–55. Springer.

Wiegand, M., Ruppenhofer, J., and Kleinbauer, T. (2019). Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Wulczyn, E., Thain, N., and Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW’17, pages 1391–1399, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Yin, W. and Zubiaga, A. (2021). Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Comput. Sci.*, 7:e598.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). 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. Association for Computational Linguistics.

Zeinert, P., Inie, N., and Derczynski, L. (2021). Annotating online misogyny. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3181–3197, Online. Association for Computational Linguistics.