UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

BRENDA SALENAVE SANTANA

A Computational-linguistic-based Approach to Support the Analysis of the Discursive Configuration of Violence on Social Media

Thesis presented in partial fulfillment
of the requirements for the degree of
Doctor of Computer Science

Advisor: Prof. Dr. Leandro Krug Wives
Coadvisor: Profª. Drª. Aline Aver Vanin

Porto Alegre
February 2023

CIP — CATALOGING-IN-PUBLICATION

Salenave Santana, Brenda

A Computational-linguistic-based Approach to Support the Analysis of the Discursive Configuration of Violence on Social Media / Brenda Salenave Santana. – Porto Alegre: PPGC da UFRGS, 2023.

150 f.: il.

Thesis (Ph.D.) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2023. Advisor: Leandro Krug Wives; Coadvisor: Aline Aver Vanin.

1. Symbolic Violence. 2. Hate Speech. 3. Intolerant Speech. 4. Frame Semantics. I. Krug Wives, Leandro. II. Aver Vanin, Aline. III. Título.

ABSTRACT

Research focused on the study of hate speech has grown in recent years; however, approaches capable of automatically detecting this type of content still need to be revised. These limitations are even more latent in languages with scarce data, such as Portuguese. This work proposes to study the use of linguistic indicators characteristic of hate speech associated with computational methods. Thus, we sought to evaluate the use of such indicators to enable the framing of content conveyed on social networks from the perspective of Frame Semantics, considering the instantiation of a *frame* of symbolic violence as a way of proposing a means of interposing discourses intolerant. As a primary source of data, we consider Twitter. From this, data was extracted covering contexts related to situations that present topics considered potential carriers of hate speech. Thus, we focus on intolerant discourses linked to the occurrence of political gender violence. In order to analyze, we created a dataset of manually annotated political context tweets. From this set of data, we were able to validate the use of the proposed frame of symbolic violence as a way of representing discourses evaluated with a higher degree of intolerance. Also, based on this data set, we performed a series of classification experiments to identify intolerant characteristics associated with hate speech. Based on this identification, we classified potential tweets with intolerant speech. To carry out the analysis of the data resulting from this study, we used a mixed method of research (quali-quanti) approach which, in its outcome, leads us to point out contributions of both scientific and social impact with which we seek to enrich the development of studies centered on social networks, with a focus on discourses potentially intolerant texts written in Brazilian Portuguese, also taking into account the user's perception of the content generated on social networks and its repercussions on daily life.

Keywords: Symbolic Violence. Hate Speech. Intolerant Speech. Frame Semantics.

Uma Abordagem Linguístico-Computacional para Auxiliar a Análise da Configuração Discursiva da Violência nas Redes Sociais

RESUMO

Pesquisas voltadas ao estudo do discurso de ódio cresceram nos últimos anos; no entanto, as abordagens capazes de detectar automaticamente esse tipo de conteúdo ainda apresentam limitações significativas. Essas limitações são ainda mais latentes em línguas com dados escassos, como a língua portuguesa. Este trabalho propõe-se ao estudo do uso de indicadores linguísticos característicos a discursos de ódio associados a métodos computacionais. Assim, buscou-se avaliar a utilização de tais indicadores para possibilitar o enquadramento dos conteúdos veiculados nas redes sociais na perspectiva da Semântica de Frames, considerando a instanciação de um *frame* de violência simbólica como forma de propor um meio de interpor discursos intolerantes. Como fonte primária de dados, consideramos o Twitter. A partir disso foi feita uma extração de dados abrangendo contextos ligados a situações as quais apresentam temas considerados potencialmente portadores de discurso de ódio. Deste modo, focamos em discursos intolerantes ligados à ocorrência violência política de gênero. De forma a analisar, criamos um dataset de tweets de contexto político manualmente anotados. A partir deste conjunto de dados, pudemos validar o uso do frame proposto de violência simbólica como uma forma de representar discursos avaliados com um maior grau de intolerância. Também a partir deste conjunto de dados realizamos uma série de experimentos de classificação com o intuito de identificar a presença de características intolerantes associadas a discursos de ódio, e a partir dessa identificação classificar potenciais tweets com discurso intolerante. Para realizar a análise dos dados resultantes deste estudo fizemos uso de uma abordagem quali-quantitativa a qual em seu desfecho nos leva apontar contribuições tanto de impacto científico quanto social, com os quais buscamos enriquecer o desenvolvimento estudos centrados em redes sociais, com foco em discursos potencialmente intolerantes escritos em português do Brasil, levando também a percepção do considerando sobre o conteúdo gerado nas redes sociais e suas repercussões no cotidiano.

Palavras-chave: Violência Simbólica, Discurso de Ódio, Discurso Intolerante, Semântica de Frames.

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
API	Application Programming Interface
CBOW	Continuous Bag of Words
CCM	Commercial Content Moderation
CL	Computational Linguistics
CNN	Convolutional Neural Network
FE	Frame Element
FN	FrameNet
GRU	Gated Recurrent Unit
GLM	Generalized Linear Model
HSD	Hate Speech Dataset
IAA	Inter-annotator agreement
LGBTQIA+	Lesbian, Gay, Bisexual, Transgender, Queer, Intersex, Asexual, and other groups and variations of sexuality and gender
LSTM	Long Short Term Memory
LU	Lexical Unit
ML	Machine Learning
NLP	Natural Language Processing
NB	Naïve Bayes
PPGC	Programa de Pós Graduação em Computação
RNN	Recurrent Neural Network
SNA	Social Network Analysis
SVM	Support Vector Machine
TCLE	Termo de Consentimento Livre e Esclarecido
TF-IDF	Term Frequency-Inverse Document Frequency

TSE	Tribunal Superior Eleitoral
UFCSPA	Universidade Federal de Ciências da Saúde de Porto Alegre
UFJF	Universidade Federal de Juiz de Fora
UFRGS	Universidade Federal do Rio Grande do Sul
UGC	User-generated Content
WN	WordNet

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

The scenario in which we currently live is overwhelmed by data (GRUS, 2015; FU et al., 2020; MATTHES et al., 2020). Websites tracking user clicks; smartphones storing location, direction, and speed every second; pedometers recording heartbeats, movements, nourishment, and even sleep patterns are examples of data usage. The Internet itself represents a vast knowledge diagram that contains an encyclopedia of cross-references (e.g., databases on films, music, sports, and politics). In recent decades, information technology has undergone an enormous evolution, with an expressive adoption of online social networks and social media platforms. The ongoing symbiotic trend regarding the increased electronic information consumption and data production by end-users using these electronic systems means that data analysis is a thriving area of continued research and development, with new resources being created daily (structured and unstructured data).

Such increasing data revolutionized communication by enabling a rapid, easy, and almost cost-less digital interaction between its users. The Internet has fundamentally changed the manner that nowadays communication takes place. One can now interact with countless individuals in different forms, changing how information is promoted and shared. Moreover, through this change in the communication paradigm between people, the Internet provided ample space for content dissemination. The spreading of opinion in a society determines the outcome of elections, the success of products, and the influence of political or social movements (BERENBRINK et al., 2022). And on the Internet, where any user can broadcast any message in these systems and reach millions of users in a short period, it is no different. The volume of user-generated content (UGC) has significantly risen. Thus, online communication has enabled information to reach people and audiences who were previously inaccessible. This democratization has been responsible for significant shifts in our culture (SILVA et al., 2016). For instance, it has opened up new channels for content production and sharing with numerous platforms and resources and is a powerful tool for diverse expression forms.

Despite the growing popularization of personal channels[1] for content dissemination, virtual interactions are still heavily dependent on text, and although its numerous advantages, the anonymity associated with it often leads to the adoption of more aggressive and hateful communication styles. According to Fortuna and Nunes (2018), on the

[1] See <https://www.tubics.com/blog/number-of-youtube-channels>

one side, people are more willing to publicly express their opinions, thereby leading to the dissemination of hate speech; On the other side, people are more likely to pursue aggressive behavior on the Web and social networks in particular because of the privacy of these ecosystems. As stated by Sousa (2019), these emerge in a fast and uncontrollable space and usually cause severe damage to their targets. In Brazil, the *Marco Civil da Internet*[2] already provides that the use of the internet in the country has as its principle the "guarantee of freedom of expression, communication, and the manifestation of thought, under the terms of the Federal Constitution," a constitution that guarantees free expression of intellectual, artistic, scientific and communication activity, regardless of censorship or license[3]. However, there is room for discussion about the limits of this freedom, often seen as unrestricted. As stated by Popper (2013), an open society must be intolerant of intolerance. A growing necessity to delineate boundaries for the right to free speech in today's global, multicultural, and multireligious societies led to the need to introduce a legal definition of hate speech (SĘKOWSKA-KOZŁOWSKA; BARANOWSKA; GLISZCZYŃSKA-GRABIAS, 2022). The impact of hate speech spans multiple areas of focus for the United Nations, from protecting human rights and preventing crime to keeping peace and achieving gender equality, and supporting children and young people. The perception of such implications led to several institutions and governments[4,5,6] becoming more seriously concerned with successfully detecting and regulating aggressive and hateful behaviors online. This type of language shares characteristics in common with another primary concern nowadays, fake news, thus reinforcing the need to mitigate such practices (CHULVI; TOSELLI; ROSSO, 2022). Even though governments are no longer the principal speech regulators (KAYE, 2019), there are still attempts to try to control speech through gatekeepers like social media platforms; however, the sheer amount of speech involved necessitates operationalizing the concept of hate speech using automated procedures and several tones of human content moderators (WILSON; LAND, 2020).

As exposed by Müller (2019a), history has shown that when ideas of understanding are lacking, individuals seek refuge in collectivities where thinking is more similar to their own. Thus, according to Nemer (2018), a specialist in Anthropology of Infor-

[2] See <https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm>

[3] See <https://www.jusbrasil.com.br/topicos/10730738/inciso-ix-do-artigo-5-da-constituicao-federal-de-1988>

[4] <https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination-0/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en>

[5] <https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml>

[6] <https://www.unesco.org/en/articles/disinformation-and-hate-speech-latin-america>

matics, the Internet provides an environment where groups with common affinities can meet. Therefore, this phenomenon provides an impulse towards the sharing of opinions that, in other circumstances, might not have so much visibility. To Dunker et al. (2018), the formation of attitudes formerly called “criticisms”, based on the productive cultivation of uncertainty, might lead to a discursive environment of the post-truth. According to the author, public opinion can buy anything, including the knowledge they agree with. Thus, the defamatory power provided by this type of speech ends up serving as a way of maintaining and reproducing prejudices. In social media and online discussion forums, user comments play a central role. Nevertheless, it is often the case where the discourse disseminated in such spaces uses different intolerant narratives to reinforce points of view and encourage other acts of violence.

Hate speech represents a form of incitement of communication that downplays or incites a person or group based on discriminatory aspects (TAY et al., 2018). The automatic detection of online hate speech has become a subject of growing interest in research in recent years (FORTUNA; NUNES, 2018). However, automatically detecting hate speech is a challenging task. As Sousa (2019) states, textual features are a limited set; they might differ from language to language. Still, there are disagreements on how hate speech should be defined. As exposed by MacAvaney et al. (2019), some content can be considered hate speech to some people, but not to others, based on their definitions. Nevertheless, existing datasets differ not only in their definition of hate speech, but lead to datasets that are not only from different sources, but also capture different information. Hence, the lack of effective mechanisms for its automated detection can most likely be due to the incorrect evaluation of toxic content (ARANGO; PÉREZ; POBLETE, 2019). Such speeches allude to the social concept of symbolic violence, which addresses a form of violence exercised by the body without physical coercion, causing moral and psychological damage (BOURDIEU, 1979). The detection of this type of speech is far from trivial due to the abstractness of the topic.

The process of identifying this type of content essentially requires the use of linguistic components. However, the linguistic field is often considered only in terms of its intersection with natural language processing (NLP) tasks (e.g., tokenization, lemmatization, stemming, part-of-speech tagging, among others), with few interdisciplinary means being applied. Hate speech detection is often approached from NLP with similar sentiment analysis techniques, i.e., identifying the opinions expressed in subjective utterances, from product and service reviews to comments on political events (AKHTAR;

BASILE; PATTI, 2019). However, online hate can be characterized by incitement to hate and violence (GLUCKSMANN, 2019), rather than merely demonstrating emotion. In that, the context in which such manifestations appear becomes essential. In the linguistic field, through semantic linguistics studies (TORRENT; ELLSWORTH, 2013; OFOGHI; YEARWOOD; GHOSH, 2006; ALAM et al., 2021; ZHENG; WANG; CHANG, 2022), relating the elements and entities associated with a given culturally incorporated human experience scene is given through frame semantics. In this representation, a frame is a schematization of experience, and a structure of knowledge, represented at the conceptual level and maintained in long-term memory (EVANS; GREEN, 2006).

Frame Semantics has been implemented computationally in the form of framenets, first for English (BAKER; FILLMORE; LOWE, 1998) and later for various other languages[7], including Brazilian Portuguese, by FrameNet Brasil (TORRENT; ELLSWORTH, 2013). Framenets can computationally express many aspects of context since they are constructed according to the concepts of Frame Semantics (TORRENT et al., 2022). The lexical choice brings a specific background frame that provides its perspective, i.e., frames provide a particular perspective (which, according to Fillmore et al. (2006), can be called a particular worldview). Within this scenario of studies, taken together with other computational approaches, using available language resources and general reference or domain-specific corpora is a key method for compiling language resources that can serve many purposes, such as detecting violence. Moreover, when considering Brazilian Portuguese, the language itself already presents a series of variations (BASSO, 2019), which influence the detection of this type of content in different ways. The particularities of the language must be considered when observing texts in that language.

Looking for the approximation of computational methods that make use of the frame semantics approach with the study of hate discourses disseminated virtually, we consider the proposal of a structure where violence is seen as a superframe that includes two sub-frames, these being a frame of physical violence and a frame of symbolic violence. Currently in Framenet Brasil there is already a frame of violence, but this is a generic one, which contemplates violence as an action that is only physical. On the contrary, we propose to make it more specific by including the notion of symbolic violence. In this study, we intend to present a formalization of linguistic heuristics proposed in the literature to automate the detection of hate speech in social networks, considering the understanding given by the semantics of frames - considering an instantiation of a frame of

[7] See <https://framenet.icsi.berkeley.edu/fndrupal/framenets_in_other_languages>

symbolic violence. Motivated by the challenges mentioned earlier, this work proposes a study on speeches spread in social networks, especially on Twitter, considering linguistic heuristics to evaluate its frame as hate speech. The choice for twitter was due to its distinction of allowing its users to view, comment, and add to all message instances or threads, unlike most social networks (LI; SUN; DATTA, 2012). Twitter provides a form of dynamic content dissemination capable of quickly giving voice and generating triggers that engage the community in different movements of social interest (ISA; HIMELBOIM, 2018; RECUERO; SOARES; GRUZD, 2020; LI et al., 2021). Thus, it is possible to view users of these social networks as citizen journalists or sensor observations (NAGARAJAN et al., 2009).

1.1 Goals

As a general goal, this thesis empirically evaluates the use of linguistic indicators characteristic of hate speech associated with computational methods, such as the classification of texts by machine learning algorithms, considering the understanding given by the semantics of frames.

Therefore, the specific objectives of this work are the followings:

- Propose a frame of symbolic violence, considering a restructuring of the frame currently available by FrameNet Brasil (in which there are no derived subframes) for an organized architecture, considering violence as a superframe derived into two subframes (physical violence and symbolic violence);

- Investigate content disseminated on social networks considering a frame of symbolic violence, presented to evaluating users, through questionnaires, aiming to fine-tune the framing of such content from this perspective;

- Explore the frame of symbolic violence, i.e., analyze texts extracted from social networks considering the schematization of conceptual structures, beliefs, and institutional practices that emerge from everyday experience, resulting in the representation of a situation, in this case, framed by the frame of symbolic violence by users;

- Experimentally evaluate the results of the classification given by annotator users to validate the proposed frame;

- Experimentally evaluate a trained computational model based on data collected by annotation/classification by users;

- Investigate means of generalizing the classification of texts according to the defined frame of symbolic violence;

1.2 Methodology

The goals of this thesis characterize it as an exploratory and descriptive research (ZOBEL, 2014; WAZLAWICK, 2020) as it aims to provide an in-depth investigation into the use of linguistic indicators characteristic of hate speech in the classification of speeches arising from social networks as such, which guides the formulation of hypotheses, at the same time that observes, registers, and analyzes the relationship between speeches considered hateful/intolerant with the proposal of a frame of symbolic violence

Due to its objectives, the thesis has both a survey and a experimental with regard to its (WAZLAWICK, 2020) procedures. The survey research was carried out through the application of questionnaires, with prior approval of the ethics committee via Plataforma Brasil[8], to collect demographic data, to a group of people adept at using social networks to identify their notes and assessments regarding potentially intolerant speeches directed at women (cis and transgender) involved in Brazilian politics, and the adequacy of the proposed frame of symbolic violence as a means of representing such instances of speech. Based on this survey, the thesis proposes using the annotated data to create a classification model to identify linguistic indicators and, then, the classification of speeches similar to those previously noted as intolerant (or not) based on the presence of such indicators. Thus, an analysis of the proposal is carried out empirically through descriptive cross-sectional observational studies. The evaluation of what is proposed in this thesis is then carried out through quantitative and qualitative analyses (ZOBEL, 2014; WAZLAWICK, 2020).

The applied nature of this thesis allows generating knowledge to help and direct future research and practical applications (WAZLAWICK, 2020) aimed at identifying potential intolerant discourses in the field of gender political violence from a frame semantic perspective. This is due to the proposal of a manually annotated dataset, as well as a classification model built from these data while analyzing the demographic data of the users

[8]CAAE no: 42861221.7.0000.5347

and the influence of this in the process of listing the intolerance of a speech and the potential representation of a frame of symbolic violence for this type of content.

1.3 Thesis Layout

In this section, we describe the organization of the remainder of the thesis. Chapter 2, overviews the main concepts related to our study. Chapter 3 explores related works described in the literature that have carried out research associated with our objectives. In Chapter 4 we present our designed proposal to automatically detect hate speech by applying a linguistic-computational approach. Chapter 5 presents a complete description of the experiments made in the development of this work and also discusses in depth the achievements, challenges, and limitations of our proposal. Finally, Chapter 6 presents the conclusions of this work and points to possible directions for future works.

Additionally, the reader can find the following information in the appendices of this work: Appendix A presents a Brazilian Portuguese version of the extended summary of the present work. Appendix B presents a Brazilian Portuguese version of the proposed symbolic violence frame. Appendix C presents the data annotation guidelines with what the data annotators that contributed to this research were presented before any interaction with the data to be annotated. Appendix D presents the Free and Informed Consent Form just as approved by the responsible ethics committee and presented to annotator users who contributed to the development of this study. Appendix E presents a sample of the questionnaire applied into data annotation process. Lastly, Appendix F presents a summary of complementary activities carried out during the doctorate period.

2 THEORETICAL BACKGROUND

This chapter introduces some concepts related to the topics addressed in the present study, which is useful for understanding its discussion. It is organized as follows: Section 2.1 introduces the Computational Linguistics area, highlighting the main concepts applied by it and its potential benefits to be addressed in the present work, with a focus on the frame semantics approach[1], in our work described mainly in terms of the FrameNet Brasil. Following this, Section 2.2 presents and discusses the concept of Symbolic Violence, defined by the sociologist Pierre Bordieu, and its relation with intersectional hate speech issues. Then, in Section 2.3, we explore some similar concepts on symbolically violent speeches; Section 2.4 elaborates on the definition of hate speech also explores the definition of the concept adopted in this work. In Section 2.5 some essential machine learning concepts are presented Finally, section 2.6 emphasizes the importance of this field of studies from the perspective of content moderation, mainly by social networks, and the potential benefits of advances in this area and its limitations.

2.1 Computational Linguistics

Language crucially occurs in thought, action, and social relations (CHOMSKY, 2017). Evans and Green (2006) argues that language allows fast and compelling expression and gives a well-developed means of encoding and transmitting complex and inconspicuous thoughts. Still, according to the authors, these ideas of encoding and transmitting turn out to be critical, as they relate to two essential capacities associated with dialect, i.e., the symbolic and the interactive functions. The first symbolizes concepts and, through them, enables symbolic assemblies by serving as prompts for the construction of much richer conceptualizations. The second encodes particular meanings, but also that, under these meanings and the forms employed to symbolize these meanings, which constitute part of shared knowledge in a specific speech community, language can serve an interactive function, facilitating and enriching communication in several ways.

Human language is handled by computer systems in each division of modern society, with different purposes. . Computational Linguistics (CL) is an interdisciplinary

[1] In the 70's, Minsky (1974) proposed the concept of *Frame* in Artificial Intelligence field. By its definition, frames are a type of data structure representing "stereotyped situations", which divides knowledge into substructures. Although there is some overlap, and thus similarity, between the concept in AI and cognitive linguistics, the latter's definitions are followed in this work.

field that applies computational approaches to analyzing, synthesizing, and comprehending written and spoken language (MITKOV, 2004). Freitas (2022) reinforces that as an applied area, computational linguistics is dedicated to the resolution of problems or tasks that are central to language - which does not mean, indeed, that, in order to solve problems (applied dimension), it is not necessary to investigate questions (theoretical dimension). According to Schubert (2020), the practical goals of the field are broad and varied, covering tasks such as machine translation; question answering; text summarization; analysis of texts or spoken language for a topic, sentiment analysis, or other psychological attributes; dialogue agents for accomplishing particular tasks; and ultimately, creation of computational systems with human-like competency in dialogue, in acquiring language, and in gaining knowledge from text. In this section, we focus on computational semantics and pragmatics branches among the different CL sub-areas.

The most notable point in natural language is its capacity to generate meaning (BURCHARDT et al., 2020). Studies in computational linguistics, primarily focusing on computational semantics, design meaningful representations and establish strategies for automatically assigning and reasoning those representations (ERK, 2018). As stated by Ruas et al. (2020), the association of words in a sentence often tells us more about the underlying semantic content of the document than its literal words, individually. Some methods build semantic representation based on the content analyzed, i.e., they focus on how text components relate and are used to formalize a model that provides a broader understanding of the content to humans

In the words of Schubert (2015), "as builders of potentially huge, complex systems dependent on symbolic representations, we also have to be concerned with the naturalness of the representations from our perspective". Intuitively, it should be easy to see if a putative semantic interpretation of a given sentence (in context) captures its semantic essence (SCHUBERT, 2015). Semantic representations range from general discourse models based on first-order logic to embedding representations of words or phrases, varying with context. Abend and Rappoport (2017) defined the semantic representation as a manner to a model "reflects the meaning of the text as a language speaker understands it". In this work, semantic representation is assumed as a model whose meaning is potentially understood. In this sense, a semantic representation should be associated with a method of extracting information from it, which can be reviewed and evaluated directly by humans. So, according to the authors, the extraction process should be reliable and computationally efficient. Different approaches to semantic representations require different semantic

schemes. Also, as stated in Abend and Rappoport (2017), semantic schemes diverge in whether they are anchored in the words and phrases of the text (e.g., all types of semantic dependencies) or not (e.g., logic-based representations).

Schemes based on semantic dependencies aim to extract the meaning of words and sentences, linking arguments to predicates once the predicate's arguments are semantically subordinate. Semantic dependencies are caught on in terms of predicates[2], and their arguments[3] (POLGUÈRE; MEL'ČUK, 2009). Identifying semantic dependencies is helpful for a range of problems, such as question answering, dialogue systems, and information extraction, to deeper understand meanings and their connections. The development of manually constructed resources has been vastly important in driving the field forward. Examples include WordNet[4], PropBank[5], FrameNet[6], and VerbNet[7]. Such resources provide human-generated data of high quality that can be used to train machine learning systems defining the linguistic structures to be addressed in automated analysis (ERK, 2018). Semantic structures are organized relative to conceptual knowledge structures (EVANS; GREEN, 2006).

To Evans and Green (2006), one proposal concerning the association of word meaning is based on the idea of a frame against which word-meanings are understood. The authors state frames as nitty-gritty information structures or patterns rising from everyday experiences. From this point of view, knowledge about the meaning of a word is, in part, knowledge of the individual frames with which a word is related (EVANS; GREEN, 2006). In the words of Lakoff, Dean and Hazen (2004), framing is about getting language that fits in a worldview. To the authors, it is not just language, i.e., the ideas are primary, and the language carries and evokes those ideas.

So far, we have discussed lexical, syntactic, and semantic analysis. However, human language has unique nuances that occur in specific contexts (SANTANA; VANIN, 2020). Fully understanding these contexts and extracting meaning from texts may thus require acquainting specific knowledge beyond the syntactic or the structural part of a text. In Huang (2017), pragmatics is broadly defined as "the study of language use in context". Jurafsky (2004) brought it to the computational field by defining computational pragmatics as the computational study of the relation between utterances and context. In

[2]In theories of syntax and grammar, the predicate of a sentence corresponds to the main verb, and potentially to any auxiliary verbs that accompany the main verb.

[3]It is an expression that helps to complete the meaning of a predicate (a verb).

[4]Available in: <https://wordnet.princeton.edu>

[5]Available in: <https://propbank.github.io>

[6]Available in: <https://framenet.icsi.berkeley.edu/fndrupal/>

[7]Available in: <https://verbs.colorado.edu/verbnet/>

this sense, a pragmatic analysis may play a fundamental role by mediating the relation between lexical representation and perceived meaning. As already stated, language provides a structure designed to express thoughts, actions, and social relations. In this way, its study should allow context-dependent utterance generation and interpretation.

From the perspective of the interactive function (EVANS; GREEN, 2006), language can be utilized to form scenes or frames of experience, indexing, and even developing a specific context (FILLMORE, 1982). In other words, Evans and Green (2006) synthesize that the language use can invoke frames that assemble rich knowledge structures, which serve to call up and fill in background knowledge. Frame Semantics theory states that the meanings of most words can be better understood based on a semantic frame, a description of a type of event, relation, or entity and the participants in it. The idea of framing a word's meaning in specific contexts is addressed, for instance, by FrameNet sharing roles across predicates that evoke the same frame type. More information about FrameNet, focusing on the Brazilian variation, is further discussed in the following section.

FrameNet is a project focused on applying the theory of Frame Semantics (FILLMORE, 1985) by developing an electronic frame-based dictionary (EVANS; GREEN, 2006). FrameNet is a lexical resource with distinguishing characteristics that differentiate it from other resources such as commercially accessible dictionaries and thesauri, as well as the most the well-known online lexical resource (RUPPENHOFER et al., 2016), WordNet. According to Esra'M (2019), WordNet differentiates between word senses and groups semantically similar words in hierarchical synonymy sets - synsets) unlike FrameNet. FrameNet is a knowledge base - FN database - valuable for NLP applications (PETRUCK; ELLSWORTH, 2018) and linguistics.

As frames are fundamentally semantic representations, they are similar across languages and, over the years, many projects[8] relying on the original FrameNet as the basis for additional non-English FrameNets have emerged (e.g., French, Spanish, Portuguese). In this work, we focus on the use of the FrameNet Brasil[9], which is supported by Universidade Federal de Juiz de Fora (UFJF). This research group has been working to create a FrameNet-style lexical database for Brazilian Portuguese. In the project reports, Salomão, Torrent and Sampaio (2013) refer that the project makes accessible to the public a fraction of the frames and lexical units annotated from a corpus of around 72 million

[8]Non-English FrameNet projects: <https://framenet.icsi.berkeley.edu/fndrupal/framenets_in_other_languages>

[9]Available in <https://www.ufjf.br/framenetbr-eng/>

tokens, covering a variety of uses exclusively in Brazilian Portuguese. In a general way, FrameNet is a corpus-based lexicographic and relational database (sort of a complex dictionary) of English frames that includes lexical units that evoke them, annotated sentences containing certain lexical units, and a hierarchy of frame-to-frame relations.

According to the FrameNet documentation, the concept of frame element (FE), within a frame structure, concerns a frame-specific defined semantic role that is the basic unit of a frame. Entities, attributes, events, and spatial and temporal circumstances are all examples of FEs in a scene (TEIXEIRA; ZAMORA, 2019). Frame Elements can be categorized according to the thematic role they perform. They can be called: core and non-core, which are further subdivided into peripheral and extra-thematic elements (RUPPENHOFER et al., 2016). Moreover, the corresponding word senses (lexical units) also evoke the frame.

As set by Teixeira and Zamora (2019), the core frame elements are those that represent an inherently conceptual aspect of the frame, distinguishing it from others. They are thus central to the frame since they specify it and are implied from it, even though they are not lexicalized (ibid.). Regarding non-core frames, there is a subdivision between peripheral and extrathematic frames. Peripheral FEs are elements with generic properties that can be applied to various frames and convey circumstantial information, such as place, time, mode, and function (TEIXEIRA; ZAMORA, 2019). They normally fill the roles of deputies. They are not exclusive to a frame, but they can be used to alter any frame of the appropriate type (OSSWALD; VALIN, 2014). As for the extratematical FEs, Fillmore (2007) states this as roles used to annotate a "word or phrase which can be thought of as introducing a new frame, rather filling out the details of the frame evoked by the head". Figure 2.1 presents a brief illustration of the *Cause_harm* frame described in FrameNet following such a structure. Depictive, and iteration are examples of extrathematic FEs that do not belong in the focus frame despite being present in a possible scene evoked related to this frame.

As stated by Fillmore (2007), the FrameNet project aims to generate valency definitions of frame-bearing lexical units (LUs) in semantic and syntactic terms, and it does so using word use assurances from a vast digital corpus. Some of the lexical units belonging to the *cause_harm* frame are also illustrated in Figure 2.1. Thus, it is possible to evoke a frame from the vocabulary used in a situation. However, it is noteworthy that perspective-taking frames are often abstract or non-lexical in that they are evoked indirectly by lexical units from one of their perspective-taking frames.

Figure 2.1: Adaptation of the *Cause_harm* FrameNet's representation.

Cause Harm

Definition: The words in this frame describe situations in which an Agent or a Cause injures a Victim. The Body_part of the Victim which is most directly affected may also be mentioned in the place of the Victim. In such cases, the Victim is often indicated as a genitive modifier of the Body_part, in which case the Victim FE is indicated on a second FE layer.

Core frame elements:

• Agent [Agt]	Agent is the person causing the Victim's injury.
• Body_part [BodP]	The Body_part identifies the location on the body where the bodily injury takes place.
• Cause [Cause]	The Cause marks expressions that indicate some non-intentional, typically non-human, force that inflicts harm on the Victim.
• Victim [Vic]	The Victim is the being or entity that is injured. If the Victim is included in the phrase indicating Body_part, the Victim FE is tagged on a second FE layer.

Non-core frame elements:

• Circumstances []	Circumstances describe the state of the world (at a particular time and place) which is specifically independent of the Cause_harm event and any of its participants.
• Concessive []	This FE signifies that the state of affairs expressed by the main clause (containing the Cause_harm event) occurs or holds, and something other than that state of affairs would be expected given the state of affairs in the concessive clause.
• Iterations [Iter]	Iterations refers to the number of times the Agent causes harm to the Victim.
⋮	⋮

Lexical Units: bash.v, batter.v, bayonet.v, beat up.v, beat.v, belt.v, biff.v, bludgeon.v, boil.v, break.v, bruise.v, buffet.v, burn.v, butt.v, ...

Source: The author (2023)

2.2 Intersectionality and Symbolic Violence

When discussing meanings and contexts, it is crucial to highlight that the same frame might refer to different issues (e.g., a frame about prejudice may refer to many types of discrimination). Research aimed to analyze how the intersectionality study area covers social issues from a diverse point of view. Intersectionality refers to a transdisciplinary theory that aims to apprehend the complexity of identities and social inequalities through

an integrated approach (BILGE, 2009). While intersectionality increased in prominence, it was perceived and debated in different ways (HANKIVSKY, 2014). In Collins and Bilge (2020), the common sense comprehension about this concept is:

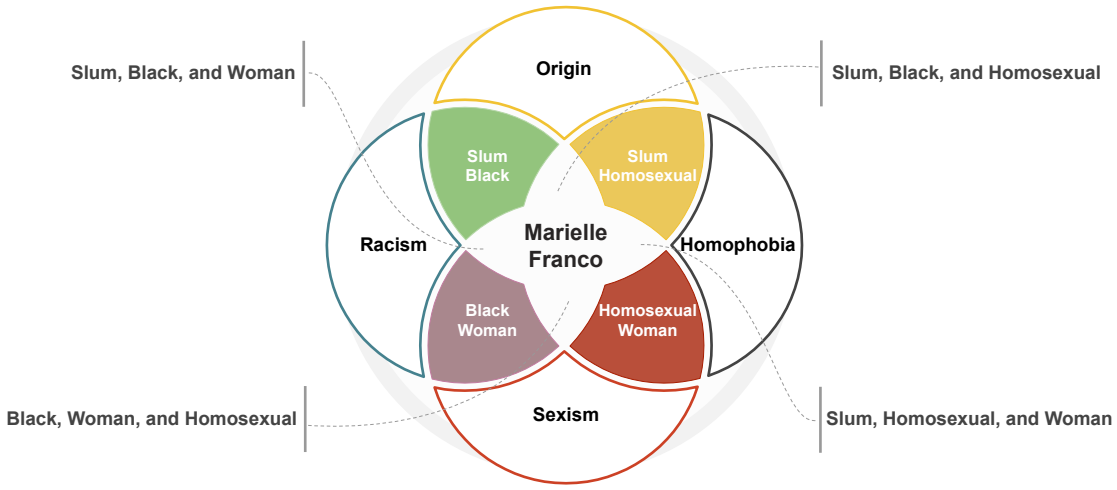
> a way of understanding and analyzing the complexity in the world, in people, and in human experiences. The events and conditions of social and political life and the self can seldom be understood as shaped by one factor. They are generally shaped by many factors in diverse and mutually influencing ways. When it comes to social inequality, people's lives and the organization of power in a given society are better understood as being shaped not by a single axis of social division, be it race or gender or class, but by many axes that work together and influence each other. Intersectionality as an analytic tool gives people better access to the complexity of the world and of themselves. (...)

The term intersectionality addresses multiple social issues that have often been set aside for a long time or discussed as isolated frames. When considering the Brazilian scenario, hatred towards minorities shows itself as an intrinsic social mark of intersectional circumstances as it reflects research made by IBGE (2019): regardless of education level, black people continue to receive much less money than white people; and situation worsens when considering the gender cut. Analyzing more Brazilian statistics, Dadico (2020) disclose that data indicate that hate overemphasizes people from groups identified by criteria of race, colur, ethnicity, sex, sexual orientation, gender identity, national and regional origin, homelessness or disability, among other attributes that expose them to greater social vulnerability. One aspect that we seek to consider in this work is the growing wave of hatred that presents itself virtually in situations that cover such issues.

Figure 2.2 illustrates a case where intersectional attacks towards an entity were spread on the Internet. The figure seeks to illustrate what happened after the 2018 murder of Brazilian sociologist and politician Marielle Franco, where a network of hatred and misinformation generated several online comments (also offline) attacking her image for several characteristics she had, and even other traits that were inadvertently attributed to her. Teixeira and Zamora (2019) highlights that Marielle - a black woman, admittedly bisexual, from a slum, a political defender of human rights - was undoubtedly crossed by all kinds of oppression triggered by the sexist, racist and classist system. Alternatively, as summarized in Brum (2019) words', "Marielle Franco welcomes in her body all minorities crushed during 500 years in Brazil". As exposed by studies that have studied the case further (SCHIRMER; DALMOLIN, 2018; BOAVENTURA; FREITAS, 2020), the rhetoric used is harmful. The attacks recorded in this case were driven by hate. Discourses related to gender, especially those with sexist content can be found not only in the political sphere but in every social context.

Violence, in this case, is imminent - including in the language used. This case

Figure 2.2: Example of Intersectionality between hate speech classes occurred in Marielle Franco's case. This example illustrates the different fronts on which speeches related to the then-councilwoman Marielle Franco were used to degrade her image. This example describes four main fronts: origin, race, sexual orientation, and sex; however, speeches related to other identity groups of which Marielle was a member were also widely disseminated after her murder.



Source: The author (2023)

exemplifies a physical violence occurrence with virtual repercussions that are also violent but, in this case, through language. Thus, through a whole symbolic production, via language and other symbolic systems, which reinforce asymmetric and hegemonic relations, disqualifications, prejudices and violence of all kinds, a symbolic violence occurs. Bourdieu et al. (1989) exposes symbolic violence as a form of violence exercised by the body without physical coercion, causing moral and psychological damage. In this sense, hate speech makes the subject invisible, and causes a death that is not so physical, but symbolic. As stated by Sardenberg (2011), symbolic violence infiltrates our whole culture and legitimizes other types of violence.

In Thapar-Björkert, Samelius and Sanghera (2016), understanding symbolic violence in contrast with conventional violence discourses is settled as essential as it provides a deeper insight into the 'workings' (i.e., complexities) of violence, provides new ways of conceptualizing violence across a variety of social fields and brings valuable intervention strategies. Following the definition provided by the Cambridge Dictionary[10], public speeches that manifest hate or encourage violence towards a person or group based char-

[10]Available in: <https://dictionary.cambridge.org/pt/dicionario/ingles-portugues/hate-speech>

acteristics as race, religion, sex, or sexual orientation, configure hate speech.

2.3 Violence in Language

Violence through the use of language can take shape in different ways, either through the use of insulting words directly addressed to the one who is seen as a target or even through speeches that are not necessarily directed at one but that are speeches that exalt practices of violence, such as speeches extolling white supremacy. Nevertheless, it should be borne in mind that, as pointed out by Butler (2021), it is not just circumstances that make words hurt. Or, as stated by the authors, we could be led to claim that all words are susceptible to being words that hurt, depending on how they are used, and that the use of words is not reducible to the circumstances of their utterance. As uttered by Toni Morrison upon receiving the Nobel Prize for Literature in 1993 [11]: "Oppressive language does more than represent violence; it is violence".

In the literature, different but similar terms can be framed as symbolically harmful speeches (e.g., dangerous speech, toxic speech, hate speech, and others.). When dealing with hate speech-related literature, different terms come up with blurred boundaries (JAHAN; OUSSALAH, 2021). However, although similar, these concepts differ from each other. Below, we present the conceptualization of the main associated topics. The order in which the definitions are presented follows our understanding of the most comprehensive to the least comprehensive concept.

1. **Dangerous Speech:** Term coined by researcher Susan Benesch, this kind of speech is any form of expression (speech, text, or images) that can increase the risk that its audience will condone or participate in violence against members of another group (MAYNARD; BENESCH, 2016).

    a. Hallmarks: Dehumanization; Accusation in a Mirror; Threat to Integrity or Purity; Assertion of Attack Against Women and Girls; and Questioning In-Group Loyalty.

2. **Toxic Speech:** Tirrell (2017) states toxic speech as mechanisms by which speech acts and discursive practices can inflict harm, making sense of claims about harms arising from speech devoid of slurs, epithets, or a narrower class that the author calls as ‘deeply derogatory terms.’

[11] See <https://www.nobelprize.org/prizes/literature/1993/morrison/lecture/>

3. **Hate Speech (a.k.a., Intolerant Speech):** Although there is no consensus on its formal definition, Fortuna and Nunes (2018) presents hate speech as "a 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 humor is used".

    a. Hallmarks: Sanction Speech; Passionate Hate and Aversion to the Different ones; and Themes and Figures of Opposition (BARROS, 2014).

As cited in these given definitions and also from the study of researches (BARROS, 2014; MAYNARD; BENESCH, 2016) some hallmarks characteristics are pointed out for those symbolically violent speeches. In the sequence, we describe these hallmarks in more detail.

1. *Dehumanization:* Speakers may convince their listeners to deny others any of the moral respect they offer to those who are "fully" human by portraying other groups of people as anything other than human, or less than human (MAYNARD; BENESCH, 2016). Dehumanizing purposes prepare followers to condone or commit violence by making their targets' death and misery seems less critical or evenuseful or required.

2. *Accusation in a Mirror:* Believing that you, your family, your community, or even your society are facing an existential threat from another group makes it seem not just reasonable (as dehumanization does) but necessary to fend off that threat.

3. *Break of Social Contracts of Integrity or Purity*

    a) Threat to group Integrity or Purity: assert that members of another group can cause irreparable damage to the integrity or purity of one's own group.

    b) Assertion of Attack Against Women and Girls: It is the suggestion that women or girls of the in-group have been or will be threatened, harassed, or defiled by members of an out-group. In many cases, the purity of women symbolizes the group's purity, identity, or way of life.

    c) Sanction Speech: Such discourses intend to punish subjects considered bad complaints of certain social contracts. Those who defend ideas out of what is socially expected are a target of these speeches;

4. *Questioning In-Group Loyalty:* In a dangerous speech, out-group or target group members are generally identified; some of it never mentions them, instead characterizing in-group members as insufficiently loyal or even treacherous to be sympathetic to the out-group.

5. *Figures of Opposition*

   a) Passionate Hate and Aversion to the Different ones: Speeches in which the passions of hatred and fear prevail in relation to what is considered different. These occur from antipathy to homophobia, racism, xenophobia, and misogyny, among others;

   b) Themes and Figures of Opposition: Speeches that develop themes and figures from the opposition between equality or identity and difference.

It is possible to notice a remarkable similarity between some hallmarks. Considering such similarities, we understand that the hallmarks can be organized into five major sets (Dehumanization; Accusation in a Mirror; Break of Social Contracts of Integrity or Purity; Questioning In-Group Loyalty; and the Construction Figures of Opposition), which include other specific hallmarks. Table 2.1 shows the suggested organization.

Table 2.1: Hallmarks Organization.

Macro Hallmarks	Specific Hallmarks	Dangerous Speech	Toxic Speech	Hate Speech
Dehumanization	Dehumanization	■	■	■
Accusation in a Mirror	Accusation in a Mirror	■	■	
Break of Social Contracts of Integrity or Purity	Threat to Integrity or Purity	■	■	
	Assertion of Attack Against Women and Girls	■		
	Sanction			■
Questioning In-Group Loyalty	Questioning In-Group Loyalty	■		
Figures of Opposition	Themes and Figures of Opposition			■
	Passionate Hate and Aversion to the Different ones			■

Due to its generalization and closer approximation to what is proposed in this work, the concept of hate speech proposed by Fortuna and Nunes (2018) was adopted as a definition of what is considered hate/intolerant speech in this work.

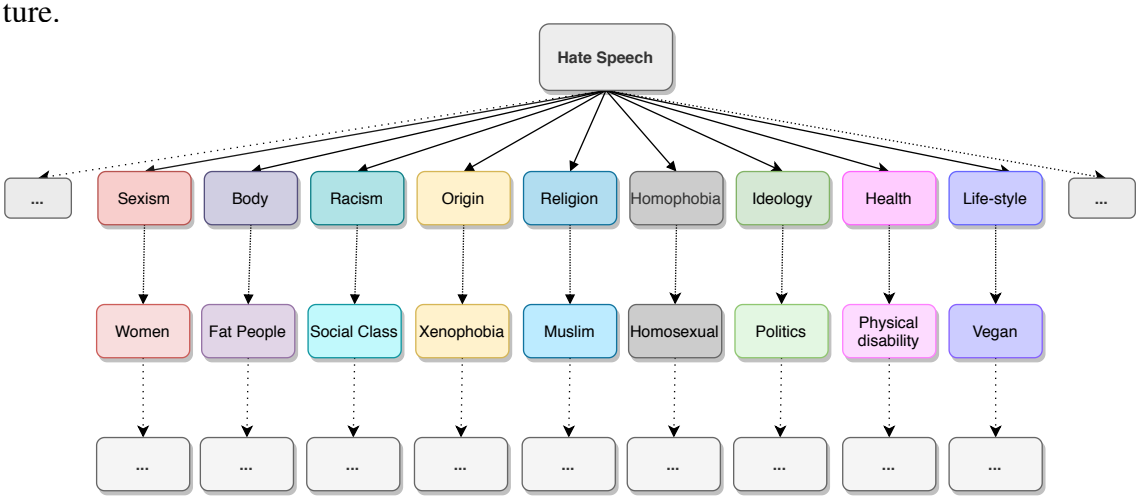
2.4 Hate Speech

Hate speech as violence reflects circumstances in which violence is often practiced in a dimension that is not physical, but they are often veiled. Therefore, when expressed with words as a form of violence, such speeches subject the target to discrimination, injury, and denial of recognition.

Hate speech does not rely based on a single identity. As pointed out in Richardson-Self (2018), choosing groups with protected characteristics covered by hate speech laws is a common feature of all definitions of hate speech. Illustrating the group characteristics that usually seem the target of such discourse, Fortuna and Nunes (2018) presents a series of hate speech classes with a directed acyclic category graph structure. An expanded version of this structure is presented in Figure 2.3. Hate speech is linked to the use of words that insult, intimidate or harass people sharing a common property, which often composes historically subjugated social markers (e.g., race, color, ethnicity, nationality, sex, gender, sexual orientation or religion, social origin, socioeconomic position, level educational status, migrant or refugee status and disability). Furthermore, new social groups may become the target of hatred over time and through space (BAIDER, 2020).

Nevertheless, it is essential to highlight cases where the hate speech category is not linked to a single class. In this type, the hate expressed may be towards more than one community and identity (e.g., Marielle Franco's case example presented in Section 2.2) - this type of hate speech is known as hybrid hate speech (CHETTY; ALATHUR, 2018). Although the usual classification of hate speech is based on characteristics such as those already presented, Baider (2020) points out that limiting hate speech to groups with protected features may fail to safeguard other vulnerable populations without state protection. Generally speaking, there is no consensually accepted definition in the literature regarding the definition of hate speech. In this sense, based on the definition adopted by each one, there may be differences in the classification of content as hate speech or not (MACAVANEY et al., 2019).

Figure 2.3: Hate speech classes represented with a directed acyclic category graph structure.



Source: Based on Fortuna and Nunes (2018).

Baider (2020) presents an understanding of hate speech as a process, where the author states three types of relationships involved: alienation, subordination, and dehumanization. First, to generate an in-group/out-group dichotomy, the alienation process includes modifying a person's perception of a group. This is achieved by highlighting (real or imagined) distinctions while ignoring or neglecting similarities. As the author points out, the hostile 'other' so created starts to be seen and classified as socially inferior or threatening over time, reifying antagonism and arousing distrust. In this scenario, Baider (2020) emphasizes that the social dynamics designed reflect these subordination relationships because of the inferior status assigned to the out-group and its members or because of the danger that their 'difference' poses to the rest of the community within the group (ibid.). In the third step of the process, the Other is then dehumanized, which is pointed to as the primary function of hate speech (WALDRON, 2012).

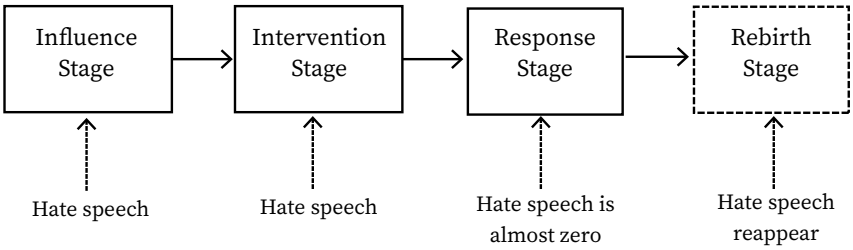
ElSherief et al. (2018) point out that there are cases where hate speech can be directed at a specific individual (Directed) or directed at a group or class of people (Generalized). In addition to being more personal and directed, their research shows that directed hate speech is more informal, angrier, and sometimes explicitly attacks the target directly (namely) with fewer analytical terms and more words implying authority and power. On the other hand, generalized hate speech is dominated by religious hate, marked by the use of deadly terms such as murder, exterminate, and kill; and quantity words such as million and many.

The dynamics of hate speech involve the occurrence of triggers that favor their manifestation (ALMEIDA; CUNHA, 2020). Such triggers are usually socially polarizing issues involving debates about elections, abortion, racial quotas, etc. Linked to these triggers, periods like elections also increase the number of denunciations of such speeches[12]. In the case of the increasing incidents of hate speech during campaign rallies and the aftermath of the election, according to (COLLINS, 2017), seemingly contributed to the re-emergence of hate speech in public venues as well as an increase in accompanying violent acts. In this same line, Chetty and Alathur (2018) point to four different stages of hate speech that might occur following a trigger event: (1) Influence Stage; (2) Intervention Stage; (3) Response Stage; and (4) Rebirth Stage. Such stages are illustrated in Figure 2.4. As stated by the authors, at the start immediately after the occurrence of a trigger event, hate speech flows heavily on social networks (stage 1); following this phase, its massive occurrence (stage 2) decreases after a few days and decreases even more (tend-

[12]See: <http://saferlab.org.br/o-que-e-discurso-de-odio/index.html>

ing to zero) after a few more days (stage 3). Based on the form and the effect of an event, after a long time, the hate speech topic might occur again (stage 4).

Figure 2.4: Stages of hate speech following a trigger event.



Source: Chetty and Alathur (2018).

Hate speech has a considerable effect and threatens the right of the offended individual to equality and freedom. Hate speech can directly or indirectly damage the victims (CHETTY; ALATHUR, 2018). In its direct form, the hate speech victims (targets) are instantly wounded by the hate speech material. While in its indirect way, the hate speech damage can be immediate or deferred, and the agents perpetrate the delayed damage, not the original actor (e.g., when a power figure spreads hate speech and its followers massively replicate it).

Chetty and Alathur (2018) highlight that across different situations, the effect of hate speech is not the same, depending on the individual involved, content, place, and circumstances. As stated by the author, the who, when, where, and the situation determines a hate speech's effect and strength. Fully understanding contexts and extracting meaning from texts may require acquiring specific knowledge beyond a text's syntactic or structural part.

Although online interactions mainly depend on text, as Sousa (2019) states, textual features represent a limited set; they might differ from language to language. Words rarely convey their literal meaning (HE et al., 2020). People often use terms that deviate from their conventionally accepted definitions to express complex and implied meanings. The speech record is different from the written record and even more from the self-record that takes place online. There is communication in the voice that escapes the text. There is always intonation, a way of speaking, an emotion in the voice, a small amount of sarcasm, and irony, which, when represented textually, often escapes existing natural language processing techniques. Regarding hate speech, there is still subjectivity, in the sense that there are diverse valid beliefs about what the correct data labels should be (RÖTTGER et al., 2021). There are disagreements on how hate speech should be defined. So, as exposed by MacAvaney et al. (2019), this means that some content can be considered hate speech

to some people but not others, based on their definitions. There are a variety of valid viewpoints on what does (not) fall under a hate speech concept (KHURANA et al., 2022). However, automatically detecting hate speech is a challenging task. As stated by Barrett et al. (2022), this is because they take various forms, change dynamically, and are found in only a small minority of relatively short texts.

Certain sensitive content authors may purposefully omit utilizing sensitive phrases to take advantage of the absence of human audits with long experience fighting hate-detection technologies (HE et al., 2020). Recognizing implicitly damaging texts is difficult owing to their extremely context-sensitive and metaphorical arousal content. Different strategies have been proposed in the literature to identify potential texts containing hate speech. Some of that focus mainly on linguistic-based approaches, often analyzing the vocabulary used (lexical analysis). Also, automatized approaches consider semantic features by applying frames to analyze the text content in its context. Other approaches focus mainly on machine-learning-based approaches, i.e., aiming to train a computational model to identify intrinsic characteristics of data previously labeled as hate speech and non-hate speech to generalize to other domains. However, many of these approaches struggle in a fight against implicit harmful texts, considered a key challenge for text classification and semantic comprehension. Unfortunately, none of the traditional text representation models have used contextual representation as direct input when encoding the current word, which is required to understand the implicit meaning (HE et al., 2020).

Even so, among existing tools for computational analysis, we highlight the Perspective API (LEES et al., 2022). The API[13] (Application Programming Interface) uses machine learning models to determine the perceived effect a comment could have on a conversation. Figure 2.5 presents an illustration of the output provided by this API. This is not a tool created specifically to detect hate speech; however, we consider its relevant contributions to identify texts with symbolically harmful potential. For example, this tool does not distinguish hate speech and offensive speech. However, a text can be analyzed considering its "toxicity" - a metric used to identify a "rude, disrespectful or unreasonable comment that is likely to make one leave a discussion". In addition to the toxicity metric, for the Portuguese language, five other metrics are also available:

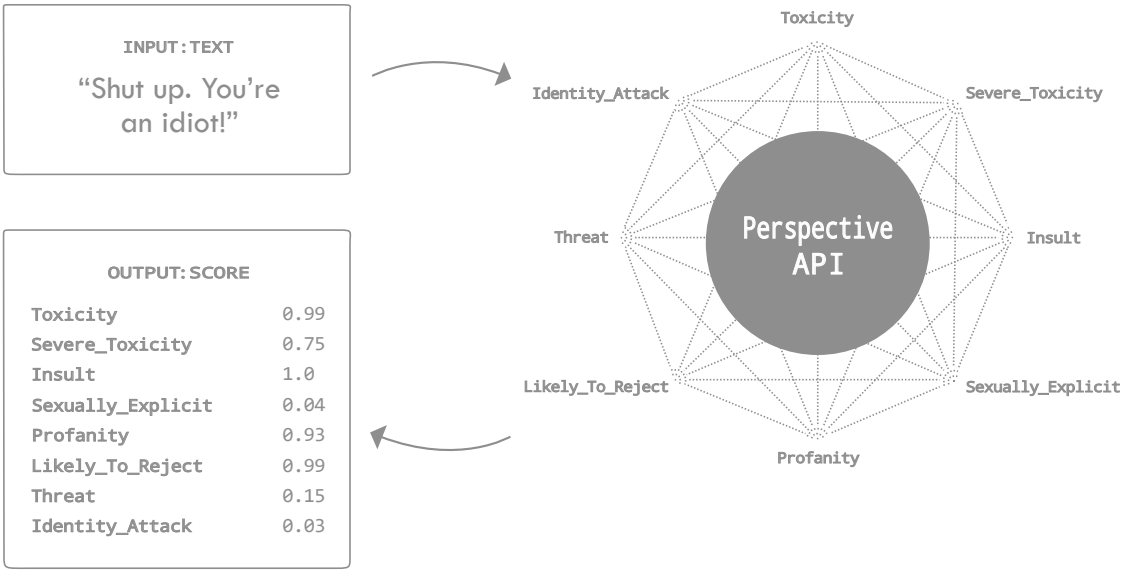
- Identity Attack: Negative or hateful comments targeting someone because of their identity;
- Insult: Insulting, inflammatory, or negative comment towards a person or a group

[13] Available in: <https://www.perspectiveapi.com>

of people;

- Profanity: Swear words, curse words, or other obscene or profane languages;
- Severe Toxicity: Severe toxicity classifies rude, disrespectful, or unreasonable comments that are very likely to make people leave a discussion;
- Threat: Describes an intention to inflict pain, injury, or violence against an individual or group.

Figure 2.5: Perspective API Example Illustration.



Source: Perspective API.
<https://support.perspectiveapi.com/s/about-the-api>

Despite Perspective API being widely used for toxicity evaluation and contributing to creating safer environments for online communication, most existing works focus on English (JIAWEN et al., 2022).

As affirmed by Akhtar, Basile and Patti (2019) usually, current work on the automated identification of different types of hate speech involves supervised learning that demands manually annotated results. Still, according to the authors, as not all annotators are equally receptive to different forms of hate speech, the intensely polarizing nature of the subjects concerned raises questions about the consistency of annotations on which these systems depend. For instance, Davidson et al. (2017) points out that it is more likely that racial and homophobic tweets are labeled as hate speech, but that sexist tweets are usually categorized as offensive. A great challenge in detecting intolerant language is the

lack of labeled datasets and the limitations of cross-linguistic methods in practical applications (AHMAD et al., 2019). However, despite the efforts, hate speech detection in Portuguese greatly lags behind English (JAHAN; OUSSALAH, 2021).

2.5 Machine Learning

Mitchell (1997) defines the [general] learning problem as a situation where:

> A computer program is said to **learn** from experience *E* with respect to some class of tasks T and performance measure *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E*.

Thus, as put by Faceli et al. (2011), in Machine Learning (ML), computers are programmed to learn from experience. To do so, they employ a principle of inference called induction, in which generic conclusions are obtained from a particular set of examples. Thus, ML involves learning hidden patterns within the data (data mining) and subsequently using the patterns to classify or predict an event related to the problem (ALPAYDIN, 2014). It is essential to state that all ML techniques are also IA approaches, even though not all artificial intelligence techniques are machine learning (ALLOGHANI et al., 2020; FREITAS, 2022). Alloghani et al. (2020) states that, in essence, machine learning algorithms are built to extract knowledge and feed it into the system for more rapid and effective process management. As settled by Freitas (2022), we know that learning is not a uniform process for people: there are the ones who learn with little, and there are those who are stubborn. Likewise, what will guide the development of a learning algorithm is being able to learn well with little and having access to quality data, that is, varied and representative of what needs to be learned (FREITAS, 2022).

As maintained by Alloghani et al. (2020), ML algorithms can be classified as either supervised or unsupervised; other authors describe other algorithms as reinforcement since they learn data and detect patterns in order to react to an environment. The difference between these two main classes is the existence of labels in the training data subset (ALLOGHANI et al., 2020). Kotsiantis et al. (2007) states that the supervised machine learning approach involves predetermined output attributes besides input attributes. The algorithms attempt to identify and categorize the predefined attribute; their accuracy and misclassification, along with other performance measures, depending on the numbers of the predetermined attribute correctly predicted or categorized or otherwise (ALLOGHANI et al., 2020). Technically, supervised learning algorithms conduct analytical tasks utilizing training data and then developing contingent functions for fresh mapping

instances of the characteristic, according to Libbrecht and Noble (2015). Conversely, unsupervised data learning involves pattern recognition without the involvement of a target attribute. That is, all the variables used in the analysis are used as inputs, and because of the approach, the techniques are suitable for clustering and association mining techniques (ALLOGHANI et al., 2020). According to Hofmann (2001), unsupervised learning algorithms are suitable for creating the labels in the data that are subsequently used to implement supervised learning tasks.

There are several challenges in applying machine learning, such as deciding the most suitable algorithm and strategy for a particular situation, learning from imbalanced data, and evaluating models. These particular issues are addressed in the following subsections. A further discussion on other challenges in machine learning can be found in Paleyes, Urma and Lawrence (2022).

2.5.1 Language Representation

Domingos (2012) states that a typical machine learning system consists of three components: representation, objective and optimization. To put it another way, Liu, Lin and Sun (2020) summarizes that, in order to build an effective machine learning system, it is first required to convert useful information from raw data into internal representations such as feature vectors. Then, by creating appropriate objective functions, optimization algorithms can be used to determine the system's optimal parameter settings. As settled by Liu, Lin and Sun (2020), the amount of valuable information that may be recovered from raw data for subsequent categorization or prediction is determined by data representation. When more meaningful information is translated from raw data to feature representations, classification or prediction performance improves. As a result, data representation is an essential component for effective machine learning (LIU; LIN; SUN, 2020).

The purpose of NLP is to create linguistic-specific programs that will allow machines to understand languages (SANTANA et al., 2023). Natural language texts are typical unstructured data, with multiple granularities, tasks, and domains, making NLP difficult to accomplish satisfactory performance (LIU; LIN; SUN, 2020). Language representation and neural language models are rapidly emerging topics that will almost certainly play a significant role in the future of NLP (SCHOMACKER; TROPMANN-FRICK, 2021). Among the different approaches of representation learning for NLP, in this work we highlight: Word2vec, TF-IDF and Count vectorize. Further information regarding

these and different approaches can be found in Schomacker and Tropmann-Frick (2021) and also in Liu, Lin and Sun (2020).

As maintained by Patel and Meehan (2021), in CountVectorizer, the approach followed is just count the number of times terms occur inside the dataset, which results in weighting in favor of the most common phrases. While in TF-IDF, short for Term Frequency-Inverse Document Frequency, on the other hand, a whole document weightage of a term is examined (PATEL; MEEHAN, 2021). Those approaches are useful for dealing with the most often used word. With the same purpose but with a different strategy, there is bag-of-words model, a popular item classification representation method (ZHANG; JIN; ZHOU, 2010). The key concept behind this method, as states Zhang, Jin and Zhou (2010), is to quantize each extracted key point into one of the visual words and then display each image using a histogram of the visual words. A clustering method (e.g., K-means) is typically employed to generate the visual words for this purpose. In the same line, there is another approach called Word2vec. Word2vec is a toolkit proposed to learn word distributed representations from large-scale corpora. The toolkit has two models, including Continuous Bag of Words (CBOW) and Skip-gram (LIU; LIN; SUN, 2020). Based on the assumption that the meaning of a word can be learned from its context, CBOW optimizes the embeddings so that they can predict a target word given its context words. Skip-gram, on the contrary, learns the embeddings that can predict the context words given a target word.

2.5.2 Predictive Models

Faceli et al. (2011) defines a predictive machine learning algorithm as a function that builds an estimator given a set of labeled examples. The label or tag takes values in a known domain. Thus, still according to the author, if this domain is a set of nominal values, there is a classification problem, also known as concept learning, and the generated estimator is a classifier. Faceli et al. (2011) also points out that different ML algorithms can find different decision boundaries. Furthermore, differences in the training set, variations in the order in which examples are presented during training, and stochastic internal processes can cause the same ML algorithm to encounter different boundaries at each new training session. There are different approaches in the literature for the elaboration of predictive models, such as probabilistic methods, search-based methods, optimization-based methods, and even multiple predictive models (FACELI et al., 2011).

Bayesian learning models are examples of models implemented using probabilistic methods. In Bayesian learning, the value of a random variable has an associated probability. Thus, Bayes' theorem is used to calculate the posterior probability of an event, given its prior probability and the likelihood of the new data. In Machine Learning, naive Bayesian classifiers are a set of simple probabilistic classifiers based on the application of Bayes' theorem with assumptions of independence between features. Naïve Bayes algorithm is a supervised learning algorithm based on Bayes theorem and used for solving classification problems (MURPHY et al., 2006). This algorithm is divided into two main steps: learning and classification. In the learning stage of the algorithm, the initial classification probabilities are calculated for each existing class, that is, the chances of an attribute being labeled in each class. Afterward, the model is built by calculating the probabilities of belonging to the class according to the previously specified training base. In this process, probabilities are also smoothed to avoid erroneous calculations. In the classification stage, the probabilities of belonging to each class are calculated, applying the previously estimated model. After calculating such feasibility, the data entry is classified within the existing possibilities. Several algorithms use Bayesian learning in their implementation (SINGH et al., 2019; WICKRAMASINGHE; KALUTARAGE, 2021). Three popular variations are Bernoulli Naïve Bayes – based on the Bernoulli Distribution and accepts only binary values, i.e., 0 or 1; Complement Naïve Bayes – instead of calculating the probability of an item belonging to a particular class, it calculates the probability of the item belonging to all the classes; and Multinomial Naïve Bayes – suitable for classification with discrete features (SINGH et al., 2019). Another very popular example of probabilistic classification methods is Logistic Regression. The main focus of logistic regression analysis is classifying individuals into different groups. Cokluk (2010), defines Logistic Regression as an analysis that enables us to estimate categorical results like group membership with the help of a group of variables.

Examples of search-based methods are decision trees. The decision tree approach is a popular data mining method for constructing prediction algorithms for a target variable or establishing classification systems based on many variables (SONG; YING, 2015). Song and Ying (2015) explains that in this approach, a population is divided into branch-like segments that form an inverted tree with a root, internal, and leaf nodes. The technique is non-parametric and can deal with massive, complex datasets effectively without imposing a sophisticated parametric framework. Another popular methodology in machine learning is Gradient Boosting. Gradient Boosting is a machine learning paradigm

that uses an ensemble of weak learners to increase model performance in terms of efficiency, accuracy, and interpretability. These models are often decision trees, aggregating their outputs for improved overall outcomes. Two outstanding gradient boosting approaches are XGBoost (CHEN; GUESTRIN, 2016) and LightGBM (KE et al., 2017). XGBoost, short for *eXtreme Gradient Boosting*, is stated by Bentéjac, Csörgő and Martínez-Muñoz (2021) as a scalable ensemble technique demonstrated to be a reliable and efficient ML challenge solver. On the other hand, LightGBM, short for *light gradient-boosting machine*, is stated by the author as an accurate model focused on providing high-speed training performance using selective sampling of high gradient instances.

2.5.3 Data Balancing

A frequent issue faced by machine learning approaches is related to an imbalance in the data. He and Garcia (2009) points out that, technically, any data collection with an unbalanced distribution of classes can be considered imbalanced. There is no agreement, or standard, concerning the exact degree of class imbalance required for a dataset to be considered truly "imbalanced." (MA; HE, 2013). However, the consensus is that unbalanced data refers to datasets with substantial, and in some cases dramatic, imbalances (HE; GARCIA, 2009). This type of imbalance is referred to as a between-class imbalance. The underlying concern with the unbalanced learning problem is that imbalanced data can significantly impair the performance of most traditional learning methods significantly (HE; GARCIA, 2009).

Among the different approaches to deal with this problem, sampling methods represent a prevalent method for dealing with imbalanced data (MA; HE, 2013). Random undersampling and random oversampling are the most fundamental sampling methods (MA; HE, 2013). Random undersampling removes majority class instances from the training data at random, and random oversampling duplicates minority class training examples at random. These two sampling strategies reduce the degree of class imbalance. However, because no new information is given, any underlying difficulties with absolute rarity are ignored.

More advanced sampling methods use some intelligence when removing or adding examples. The synthetic minority oversampling technique (SMOTE) is an approach that oversamples the data by introducing new, non-replicated minority class examples from the line segments that join the five minority class nearest neighbors. For undersampling,

a more robust approach can be achieved by NearMiss algorithm (MANI; ZHANG, 2003). It aims to balance class distribution by randomly eliminating majority class examples. When two different classes are very close to each other, it is possible to remove the majority class instances to increase the spaces between the two classes. This helps in the classification process.

2.5.4 Evaluation Metrics

In several cases, only one set with n objects must be used in predictor induction and its evaluation. Therefore, alternative sampling methods should be used to obtain more reliable predictive performance estimates, defining training and test subsets (FACELI et al., 2011). The training data are employed in the induction and model fitting. In contrast, the test examples simulate the presentation of new objects to the predicted ones, which were not seen in the induction. These subsets are disjoint to ensure that performance measures are derived from a different set of examples than the one used in learning. The four main existing sampling methods are (FACELI et al., 2011): (1) holdout, random sampling, bootstrap, and cross-validation. In the case of (1) holdout, the data set is divided into a proportion of p for training and $(1 - p)$ for testing. Normally, $p = \frac{2}{3}$ is used. This approach is quite dependent on the generated partitions. Thus, to make the results less dependent on the partition made, (2) it is possible to make several random partitions and obtain an average performance in a holdout; this method is known as random subsampling. In the bootstrap method (3), *r* training sets are generated from the original sample set. Examples are randomly sampled from this set, with replacement, i.e., an example may be present in a given training subset more than once. Unselected examples make up the test subsets. The final result is then given by the average performance observed in each test subset. In the cross-validation method k-fold, the set of examples is divided into k subsets of approximately equal size. Objects from $k - 1$ partitions are used in training a predictor, which is then tested on the remaining partition. This process is then repeated k times, using a different test partition in each cycle. The final performance of the predictor is given by the average of the performances observed on each test set. A variation of this method for classification problems is the k-fold stratified cross-validation, which keeps in each partition the proportion of examples of each class similar to the proportion contained in the complete dataset. Such a technique is used in problems where the goal is prediction so that one seeks to estimate how accurate this model is to a new set of data. In a study

carried out by Rodriguez, Perez and Lozano (2010), after performing a sensitivity analysis of the cross-validation k-fold in estimating the error of predictions, it is recommended to use $k = 5$ or $k = 10$, given the low bias presented using such values.

Once the sampling methods are defined for more reliable predictive performance estimates, as described by Faceli et al. (2011), it is then possible to measure performance measures of the models. For simplicity, we consider a problem with two classes, where usually, one class is denoted as positive and the other negative. This way, it is possible to structure an evaluation matrix, popularly known as a confusion matrix. Table 2.6 illustrates such a structure. In this matrix: the true Positive (TP) corresponds to the number of examples of the positive class correctly classified; analogously, the True Negative (TN) corresponds to the number of examples of the negative class correctly classified; The False Positive (FP) corresponds to the number of examples whose actual class is negative but which were incorrectly classified as belonging to the positive class; thus, analogously to False Negatives (FN), they correspond to the number of examples originally belonging to the positive class that was incorrectly predicted as belonging to the negative class.

Figure 2.6: Confusion Matrix

		Prediction outcome		
		p	**n**	**total**
actual value	**p′**	True Positive	False Negative	P′
	n′	False Positive	True Negative	N′
	total	P	N	

As defined by Faceli et al. (2011), a series of other performance measures can be derived from this confusion matrix. Among them are (MONARD; BARANAUSKAS, 2003):

- Error rate in the positive class: proportion of examples of the positive class incorrectly classified by the predictor $\hat{f}$, also known as the false negative rate. This rate can be achieved by using Equation 2.1.

$$err_{+}(\hat{f}) = \frac{FN}{TP + FN} \tag{2.1}$$

- Negative class error rate: proportion of negative class examples incorrectly classified by the $\hat{f}$ predictor, a.k.a., false positive rate. This rate can be achieved by using Equation 2.2.

$$err_{-}(\hat{f}) = \frac{FP}{FP + TN} \tag{2.2}$$

- Total error rate: given by the sum of the secondary diagonal values of the matrix, divided by the sum of the values of all matrix elements. This rate can be achieved by using Equation 2.3.

$$err(\hat{f}) = \frac{FP + FN}{n} \tag{2.3}$$

- Total hit rate (accuracy): calculated by the sum of the main diagonal values of the matrix, divided by the sum of the values of all elements of the matrix. This rate can be achieved by using Equation 2.4.

$$acc(\hat{f}) = \frac{FP + FN}{n} \tag{2.4}$$

- Precision: proportion of correctly classified positive examples among all those predicted as positive by $\hat{f}$. This rate can be achieved by using Equation 2.5.

$$prec(\hat{f}) = \frac{TP}{TP + FP} \tag{2.5}$$

- Sensitivity or recall: corresponds to the hit rate in the positive class. It is also known as the true positive rate (TPR). This rate can be achieved by using Equation 2.6.

$$sens(\hat{f}) = rec(\hat{f}) = TVP\hat{f} = \frac{TP}{TP + FN} \tag{2.6}$$

- Specificity: corresponds to the hit rate in the negative class. This rate can be achieved by using Equation 2.7.

$$esp(\hat{f}) = \frac{TP}{TP + FP} = 1 - RTP \tag{2.7}$$

Faceli et al. (2011) states that precision can be seen as a measure of model accuracy, while recall can be seen as a measure of its completeness. However, when observing the accuracy of a model, as exposed by the author, it is possible to identify the rate of items labeled as belonging to a certain class that belongs to that class. However, it does not provide information regarding the number of examples of that class that were not classified correctly. While looking at recall allows extracting the rate of instances labeled in a

given class, but says nothing about how many other instances were incorrectly classified as belonging to this class. Thus, as stated in Faceli et al. (2011), precision and recall are generally not discussed separately but are combined into a single measure, such as the F-measure. This metric is achieved by calculating the weighted harmonic mean of precision and recall. The F measure attempts to address the convenience brought on by a single metric versus a pair of metrics (JAPKOWICZ; SHAH, 2011). For any $\alpha \in$ R, $\alpha > 0$, a general formulation of the F measure can be given as in Equation 2.8.

$$F_\alpha = \frac{1 + \alpha[Prec(f) \times Rec(f)]}{\{[\alpha \times Prec(f) + Rec(f)]\}}. \tag{2.8}$$

There are several variations of the F measure. For instance, the balanced F measure weights the recall and precision of the classifier evenly, i.e., $\alpha = 1$, a.k.a., F1-measure, illustrated in the equation 2.9.

$$F_1 = \frac{2[Prec(f) \times Rec(f)]}{[Prec(f) + Rec(f)]}. \tag{2.9}$$

2.6 Content Moderation

In order to get feedback, engage their readers, and build customer loyalty, news portals, and blogs often also allow their readers to comment (PAVLOPOULOS; MALAKASIOTIS; ANDROUTSOPOULOS, 2017). Social network platforms (e.g., Facebook, Twitter, Instagram, and Linked-in) serve an increasingly important political role as outlets for discourse. Collectively, they provide space for public members to meet, explore, debate, and exchange knowledge, a position other channels promote through their rhetoric. However, they are often cases where the shared content violates the norms of such networks. In today's global, multicultural, and multi-religious society, exercising one's freedom of speech may violate the dignity, freedom of thought, conscience, and religion of others or breach laws against discrimination (SĘKOWSKA-KOZŁOWSKA; BARANOWSKA; GLISZCZYŃSKA-GRABIAS, 2022). In the online communication context, it is where content moderation policies take place. Content moderation consists of filtering user-generated content posted on blogs, social media, and other online outlets to determine the suitability of the content for a particular platform, location, or jurisdiction (ROBERTS, 2017). As pointed out by ROBERTS (2017), content moderation can be performed by volunteers or, progressively, by individuals or companies earning remuneration for their

services in a commercial context - this latter practice is called Commercial Content Moderation, or CCM. Companies that own social media sites and networks demanding UGC use content moderation to protect the company from liability, advertisement, and curating and regulating user experience.

Content moderation is typically implemented as an AI-human hybrid process (JIANG; ROBERTSON; WILSON, 2020), and its strategy may trigger a moderator to remove UGC, acting as an agent of the platform or site in focus (ROBERTS, 2017). The rigor of moderation may vary from website to website and platform to platform. Rules around what UGC is allowed are often set at a site or platform level and reflect that platform's brand and reputation, its tolerance for risk, and the type of user engagement it wishes to attract (ROBERTS, 2017). To some degree, no platform does not impose rules (GILLESPIE, 2018). However, with minimum rigor, some social networks are known for applying a completely free speech policy, for example, the case of the social media platform Parler, which markets itself as "the world's premier free speech platform". Platforms like this present themselves as an alternative space for exchanging content and information with other networks (e.g., Facebook and Twitter) that have been growing their look at the violation of rights and attacks spread through the content found on their platforms. Nevertheless, as stated by Gillespie (2018), despite a truly "open" platform's fantasy being powerful, resonating with the deep, utopian community and democracy concepts, it is just that, a fantasy.

Since content moderation approaches might present bias (e.g., political (JIANG; ROBERTSON; WILSON, 2020)), approaches that aim to filter potentially harmful contents might be cautious but assertive in their design. Freedom of expression is neither an absolute nor an unlimited right. It is unrealistic to assume that entirely automated moderation would be flawless since that comments - or other online textual interactions - might contain irony, sarcasm, harassment without profanity, etc., which are especially difficult for computers to manage (PAVLOPOULOS; MALAKASIOTIS; ANDROUTSOPOULOS, 2017). In this sense, in our work, we stand for an approach capable of identifying and pointing out hate speech characteristics, but without imposing measures to be taken - these being in charge of practical applications for future use.

2.7 Chapter Summary

This chapter described the main concepts that permeate our goals in this work. In this way, we seek to bring a review of concepts, research, and discussions on topics addressed in this work and used in the construction of its theoretical and scientific basis.

As discussed in this chapter, violence can occur in different ways, not limited to physical aggression, including symbolic ways. One way of manifesting violence in its symbolic form is through language. Violence manifested through words can sometimes be characterized as hate speech, although there is much debate about the most appropriate definition of this concept. Although there is a broad debate about the definitions, there are hallmarks in the literature, treated in this work as linguistic indicators, which help in the characterization of these speeches (such as *Sanction Speech*; *Passionate Hate* and *Aversion to the Different Ones*; and *Themes and Figures of Opposition*, which are indicators considered in this work), as well as other related concepts (e.g., dangerous speech and toxic speech). As put by Butler (2021), it is indisputable that words hurt and irrefutably correct that hateful, racist, misogynistic, and homophobic speech must be vehemently opposed. It should be considered that such discourses can occur in transversal ways, intolerantly covering their targets in an intersectional way, that is, attacking for different characteristics that it carries in themselves.

Hate speeches are disseminated in different spheres and target different groups, and the way to attack those seen as targets is different in each context. Thus, the analysis of the context in which such situations occur are fundamental. From a linguistic point of view, language usage may trigger frames that put together complex knowledge structures and act as a call to action for filling in background information. From this, through the idea of Frame Semantics theory, computationally represented through framenets, the idea of framing a word meaning in specific contexts is addressed. Context is a key component of the Frame Semantics theory. Moreover, as stated by Torrent et al. (2022), frames can represent the immediate context against which meaning is to be interpreted. However, as far as we know, little has been explored in the literature to analyze hateful content from this perspective.

Different approaches have been proposed to contain the manifestation of hate speech in social networks. One of them implements content moderation policies, and this practice is quite dependent on humans for content analysis. Even though every social media platform does content moderation, alternative social media have recently proposed

a "free speech" environment, where milder interpretations of hateful content are implemented. Thus, it is possible to find in these networks a concentration of users with a propensity to disseminate hate – the work of Israeli and Tsur (2022) points out that hate mongers make about 16% of Parler active users, and that these users have distinct characteristics compared to other user groups.

3 RELATED WORKS

In this Chapter, related works which have carried out research analogous to the objectives of the present work are described. In Section 3.1, we explore some works that bring to the debate multiple cases where women became targets of damaging speeches in the Brazilian politics scenario through symbolic violent speech actions. After establishing the discussion of how discursive practices analogous to hate speech have been growing in this scenario, we discuss in the following sections different approaches to identifying and moderating this type of speech - not necessarily focusing on hate speech motivated by gender. To better organize the different approaches found in the literature, we present the most similar works divided into two main strands: (1) approaches mainly based on machine learning techniques - presented in Section 3.2; and (2) approaches that make use of linguistic features essentially - presented in Section 3.3. Despite the limited data available for the study of the detection of hate speech in Portuguese, section 3.4 presents a brief description of the main datasets and auxiliary resources for hate speech research in Portuguese that can be found. Finally, Section 3.5 brings a summary of the key ideas and proposals presented throughout the Chapter, reinforcing the differences and similarities from this current research.

3.1 Symbolic Violence and Women in Brazilian Politics

Throughout the twentieth century, many studies have been conducted on the issue of gender. Since then, the perception of gender as a social construct resulting from factors other than biological sex has been discussed more deeply in different areas. It was possible to identify how the distinction between male and female was explicitly linked to the imposition of subordination to women by recognizing gender as a construct (PINHO, 2020). Gender is not only a view of sexual distinctions but their hierarchy, according to Scott (1986). Moreover, this hierarchical characterization, as interpreted by Pinho (2020), which gives men the role of command and women the position of submission, also creates particular types of violence against women, which lies in the difficulty of understanding this phenomenon without taking into account the association between their motivation and the gender of the victim. Among these types of violence, there is political gender violence.

Anchored in the definition of symbolic violence previously presented in this work

(See Section 2.2), which is conceived as a practice used against others to confirm their position in the social hierarchy, in politics, this is used as a form of delegitimization through gender stereotypes that deny what is feminine competence in the political sphere. This, in the understanding of Pinho (2020), becomes violence when "it implies fundamental disrespect for human dignities, such as producing and distributing highly sexualized and pejorative images, using social media to incite violent acts, or not explicitly recognizing or denying the existence of a woman in political spaces for the simple fact of being a woman". This type of violence can be associated with one of the causes of the underrepresentation[1,2] of the Brazilian female population in national politics. By observing recent cases in the political sphere, it is possible to identify cases that illustrate the development of the growing explicit wave of disseminating injuries manifested to diverse women candidates in the 2020 Brazilian elections. These data are pointed out by a survey carried out by Tribunal Superior Eleitoral (TSE)[3], and yet another survey carried out by the Marielle Franco Institute[4], which mainly took into account attacks suffered by black women.

Under similar optics, Silva (2019) makes a deep analysis of online comments referring to the so Brazilian president Dilma Rousseff on a Facebook page in a period close to the final decision of the impeachment process (three weeks before and one after). According to Silva (2019), all the comments revolved around four central adjectives: crazy (*louca*), dumb (*burra*), whore (*prostituta*), and disgusting (*nojenta*). With the emphasis on the perception that violence begins in language, each of these attributions reveals a direct association between Internet users' opinions and the reality of the exclusion of women from public spaces (ibid.). Thus, according to the author, they demonstrate a sexist discourse reiteration, which implies discriminating, stereotyping, and marking as abnormal and inadequate female subjects who exhibit gender behaviors, conducts, and experiences that run away from what is socially determined. As noted by the author, the analysis of the statements in the comments refers to the symbolic violence to which women are generally subjected socially.

Analyzing the repercussions of the occurrence of verbal violence involving the then federal deputy Jair Bolsonaro (PSL-RJ, at the time) insulting Maria do Rosário (PT-

[1]<https://educa.ibge.gov.br/jovens/materias-especiais/materias-especiais/20453-estatisticas-de-genero-indicadores-sociais-das-mulheres-no-brasil.html#subtitulo-4>

[2]ONU Women in Politics 2020 Report: <https://www.unwomen.org/-/media/headquarters/attachments/sections/library/publications/2020/women-in-politics-map-2020-en.pdf?la=en&vs=827>

[3]<https://www12.senado.leg.br/noticias/audios/2020/11/tse-aponta-crescimento-na-violencia-contra-candidatos-nas-eleicoes-de-2020>

[4]<https://www.violenciapolitica.org>

RS), also a federal deputy, in a Brazilian plenary session held in honor of the Human Rights Day in 2014, Bittencourt and Fonseca-Silva (2020) make use of the theoretical-methodological framework of Discourse Analysis as a way to identify the effects of meaning produced in the relationship between political and legal discourses on verbal violence in the public sphere. Such work was performed by observing a collection of news related to the case, from its occurrence on December 9, 2014, to its outcome in 2019, confirming the conviction for moral damages in the Supreme Federal Court of the then deputy. The insights of the authors suggest that there is a conflict at the intersection of fact and memory between the effects of structuring and reforming verbal violence in various discursive roles in different social places that, on the one hand, generate effects such as moral harm and the violation of parliamentary decorum and, on the other hand, the effect of a direct speech.

As demonstrated by explaining outstanding cases of recent Brazilian politics, the symbolic violence demarcated through speech acts, personally and in the virtual environment, brings democracy to a chronic problem. When considering the gender cut, as exposed by a recent study developed by Terra de Direitos e a Justiça Global (2020), at least one case of an attack on life against representatives of elected offices, candidates, or pre-candidates in Brazil is recorded every 13 days. According to the same report, considering cases where the victim's gender was identified, women represent 76% of the targets of political violence cases in Brazil. Against them, the attacks have contours, such as their challenge as political agents (ibid.) to be seen as an authority. Also, according to the study Terra de Direitos e a Justiça Global (2020), the offensive and discriminatory acts mapped are based mainly on issues involving misogyny, racism, religious intolerance/racism, and LGBTQIA+ phobia. It is noteworthy that the offenses based on misogyny and racism have as the primary target black women policies.

The discursive configuration of violence permeates the central focus of this work, so in the following sections, we discuss practical approaches to identifying such cases.

3.2 Machine Learning-based Approaches

Several attempts have been made to detect hate speech on social media automatically. As stated by Fortuna and Nunes (2018), such growing interest is not just due to the increased media coverage but also the rising political interest. Allied with it, issues like the lack of automatic techniques, and the lack of data about hate speech, are pointed out

by the authors as latent issues in motivating research in this area.

Comparing seven distinct models for automatic hate speech detection (two character-based models, i.e., using character n-grams as features; and the others word-based, i.e., using word n-grams or embeddings), Gröndahl et al. (2018) argue that model architecture is less important than the type of data and labeling criteria for successful hate speech detection. It shows that when tested with any other dataset, none of the pre-trained models presents a satisfactory performance - the authors (ibid.) emphasize that across different datasets, i.e., the characteristics indicating hate speech are inconsistent. However, it is also shown (ibid.) that all models perform equally well if re-trained from another dataset with the training set and tested using the test set from the same dataset - which indicates that the detection of hate speech is mostly independent of the model architecture. According to Gröndahl et al. (2018), all these techniques are brittle against adversaries who can intentionally manipulate the text, inserting typos, changing word boundaries, or even adding innocuous words to the original hate speech models. Nevertheless, combining these methods can be effective against Google Perspective API.

Due to the high data type dependent issues, researchers (e.g., Gao and Huang (2018); Tay et al. (2018); Wilson and Land (2020)) have also been using context-aware approaches where not only the text content is observed. This approach is crucial to environments like social media, where language nuances vary. For instance, in Gao and Huang (2018), the authors propose a ML approach trained using extracted context-related features. For this purpose, the authors presented a dataset containing comments made by users on a newspaper page, as well as keeping the original news related to the topic under discussion (which, in this case, represents the context). Exploring regression models based on features and models based on neural networks, the evaluation of the approach reveals (ibid.) that context-aware logistic regression models and neural network models outperform their counterparts who are oblivious to context data. Because hate speech constantly deals with language nuances, we highlight approaches like the one proposed by Tay et al. (2018), which focuses on identifying sarcasm between textual data disseminated in social networks. In this work, the authors make use of an attention-based neural model.

Proposing a content moderation that makes use of deep learning (Recurrent Neural Network, a.k.a, RNN; and Convolutional Neural Network, a.k.a, CNN) based on approach, Pavlopoulos, Malakasiotis and Androutsopoulos (2017) demonstrate that a Gated Recurrent Unit (GRU) RNN working on word embedding outperform the previous known

state of the art - which used to apply logistic regression or multi-layered perceptron classifiers with character or word n-gram characteristics. This approach has been used to improve achievements in detecting hate speech and concomitantly in content moderation tasks.

Considering researches on content moderation linked to hate speech detection approaches, the number of studies focused on the Portuguese language is still really scarce compared with other languages with much more resources (e.g., English). However, this study field has been recently growing and showing promising results. When observing late publication using Portuguese data, many of the works (CASTRO, 2019; PAIVA; SILVA; MOURA, 2020) make use of traditional classification methods such as variations of the Naive Bayes algorithm, and Support Vector Machine (SVM). Nevertheless, different proposals which use more complex techniques and with more significant potential for the performance of this task have been emerging.

For instance, in Silva and Serapiao (2018), the authors combine a convolutional neural network to pre-trained (Wang2Vec and GloVe) and trainable word embeddings for hate speech detection in Portuguese. In this study, testing distinct optimizing functions allowed the authors to verify and reinforce the sensitivity of the model. On top of OffComBr 2 and 3 datasets[5], an F1-score higher than 89% was achieved, while 96% of the F1-score was reached in the HSD. In all cases tested, a binary classification was used (ibid.). Bispo (2018) explores using a deep cross-lingual Long Short-Term Memory (LSTM) model. According to the author, such a proposal was trained with a hate speech dataset translated from English in two different ways, preprocessed and vectorized with varied strategies that were represented in 24 scenarios (undergoing embeddings training through word index vectors, TF-IDF vectors, N- vectors Grams, with or without GloVe vocabulary), and tested with a dataset built and labeled in this work and with HSD. On top of such datasets, a precision of up to 70% was achieved in the experiments using the model trained with the corpus in English and the dataset translated into this language (ibid.). Metrics such as F1-score and accuracy were not reported. Studies like these show the potential of recent natural language processing advances that significantly contribute to the automated detection of hate speech in Portuguese.

[5] Available in <https://github.com/rogersdepelle/OffComBR>

3.3 Linguistic-based Approaches

Gröndahl et al. (2018) highlight the obstacle encountered in using more "gentle" words, which can be purposely inserted to avoid sanctions tools for automatically detecting the nature of a text, as demonstrated in his work by using the Perspective API. When considering the main focus on linguistic approaches, most hate speech detection approaches in literature use strategies based on the lexicon used in such speeches.

Davidson et al. (2017), for instance, starting from a lexicon-based search on Twitter, using a hate speech lexicon containing words and phrases identified by internet users as hate speech compiled by Hatebase.org[6] (later detailed in Section 3.4) the authors collected 85.4 million tweets containing such lexicon terms from 33,458 Twitter users. After taking a random sample of 25k tweets from this original set, a manual classification performed by CrowdFlower was applied. The data were labeled into three categories: hate speech, offensive language, or neither. After a data cleaning process over the sample data, 24.802 labeled tweets were kept, and from these, 5% were classified as hate speech by the coder's majority, and 1.3% were coded unanimously. The authors then trained a logistic regression with the L2 regularization model to differentiate between them and then analyzed the results to understand better how one can distinguish between them. Analyzing the achieved results, Davidson et al. (2017) claims that specific terms are beneficial for distinguishing between hate speech and offensive language. Nevertheless, lexical methods effectively identify potentially offensive terms but are inaccurate at identifying hate speech (ibid.); even if, according to the authors' perspective, the results also highlight the use of fine-grained labels as a helpful strategy in hate speech detection.

In Esra'M (2019), the author proposed a hierarchical domain-specific language resource of violence supported by the use combination of FrameNet 1.7 and WordNet (WN) 3.1[7] as a way to explore the lexicon of the language of physical violence scenes. Despite focusing on physical violence instead of verbal violence, the proposed approach demonstrates the potential achievements when using a computational lexicon-based approach - an F-score of 83.7% as achieved on top of a corpus representing posts and comments retrieved from Donald Trump's Facebook public page. With such a proposal, the authors highlight that the development of new frames is influenced by WN to FN, encouraging minor improvements to existing ones and supporting promising mapping between specific

[6]Available in: <https://hatebase.org>
[7]See: <https://wordnet.princeton.edu>

frames and synonymy sets (a.k.a., synsets).

3.4 Datasets and Auxiliary Resources

Few works related to hate speech detection address the Portuguese language and those that do, usually generate their own database (PAIVA; SILVA; MOURA, 2019). In this section, we describe some of the main Brazilian Portuguese datasets found in literature.

A hate speech dataset, identified by the HSD[8] acronym was suggested by Fortuna (2017), as a hierarchically annotated dataset consisting of Portuguese-language tweets collected from the Twitter site by various approaches: (i) considering unique user profiles and (ii) keywords. To extract data from such selected profiles, the author selected profiles known for posting offensive tweets on various subjects have been mentioned in. Such collection was done by searching for the keywords “hate”, “hate speech” or “offensive”. With the keywords approach (ii), were used keywords related to hate speech commonly listed in the literature obtaining hashtags, profiles and other keywords related to the topic. At the end of the process, 42,390 tweets were collected, which after the preprocessing was reduced to 5,668. Two human judges annotated this dataset. 1,228 of the overall tweets that make up the HSD, i.e., 22% of the data collection, are categorized as hate speech. The HSD was annotated with hate groups (sexism, bigotry, racism, among others), however a binary classification was adopted for all the tweets in this data collection, passing to consider either “offensive” or “non-offensive” rather than classifying in subtypes of hate speech.

Another known Brazilian Portuguese datasets were proposed by Pelle and Moreira (2017), and contain offensive (and non-offensive) comments extracted from a Brazilian news portal. The process of commenting annotation was done through three human judges, and it generated two sets: OffComBr-2 and OffComBr-3. Originally, 10,366 comments were collected; however, the authors limited it to 1,250 random samples. Those comments were then categorized into seven classes: “racism”, “sexism”, “homophobia”, “xenophobia”, “religious intolerance”, “cursing” and “non-offensive”. While this multiclass labeling was carried out by the authors, the format of the data set made available by them has binary labeling, defining it only as “offensive” and ‘non-offensive”. OffComBr-2 was created from this review, containing 1,250 comments that were noted by at least two

[8] Available in: <https://github.com/paulafortuna/master_thesis_hate_speech_portuguese>

judges as offensive or non-offensive. Although there are 1,033 findings in the OffComBr-3, the three judges have acknowledged them. In all, OffComBr-2 has 419 comments identified as offensive, 33.5% of its total comments, and OffComBr-3 has 202 comments, 19.5% of the total, identified as offensive.

Another important resource available is the already mentioned Hurlex. In Bassignana, Basile and Patti (2018), for instance, is presented a multilingual lexicon of hate words called HurtLex[9]. This lexical database had its start from an Italian preexisting lexical resource released by Mauro (2016) containing more than 1,000 Italian words, organized in 3 macro categories: (1) derogatory words, i.e., all those words that have an offensive and negative value (e.g., slurs); (2) words with prejudices, i.e., words that are typically harming people or groups belonging to marginalized categories; and (3) words that are supposed to be neutral, but context can lead to a semantic shift, turning their meaning into a negative attribute (such as when they are metaphorical). The full existent organization divided into major e finer-grained categories is present in Table 3.1. Making use of MultiWordNet[10] associated with BabelNet[11] to expand its lexicon to English, Bassignana, Basile and Patti (2018) implement a machine-readable version of the hate words lexicon manually reviewed that supports the identification of such type of speech based on its vocabulary. Currently, this lexicon is available in more than 50 languages, including Portuguese.

In a similar strategy, there is Hatebase, a set of data organized by a Canadian company. Hatebase uses a broad multilingual vocabulary based on nationality, ethnicity, religion, gender, sexual discrimination, disability, and class to monitor hate speech incidents across 178 countries. At this approach, a natural language engine, Hatebrain, performs linguistic analysis on public conversations to derive a probability of a hateful context. According to the company, all data is made available through the Hatebase web interface and API. However, only a minimal set of information about it is made available on their webpage.

[9] Available in: <http://hatespeech.di.unito.it/resources.html>

[10] A multilingual variation of WordNet, an English lexical database.

[11] It is a combination of a multilingual encyclopedic dictionary and a semantic network which, in an extensive network of semantic relationships, links concepts and named entities.

Table 3.1: HurLex Categories.

Macro Categories	**Finer-grained Categories**	**Description**
Negative Stereotypes	PS	Ethnic Slurs
	RCI	Location and Demonyms
	PA	Profession and Occupation
	DDP	Physical Disabilities and Diversity
	DDF	Cognitive Disabilities and Diversity
	DMC	Moral Behavior and Defect
	IS	Words Related to Social and Economic antage
Hate Words and Slurs Beyond Stereotypes	OR	Words Related to Plants
	AN	Words Related to Animals
	ASM	Words Related to Male Genitalia
	ASF	Words Related to Female Genitalia
	PR	Words Related to Prostitution
	OM	Words Related to Homosexuality
Other Words and Insults	QAS	Descriptive Words with Potential Negative Connotations
	CDS	Derrogatory Words
	RE	Felonies and Words Related to Crime and Imoral Behavior
	SVP	Words Related to the Seven Deadly Sins of the Christian Tradition

Source: Bassignana, Basile and Patti (2018).

3.5 Chapter Summary

This chapter briefly described some of the most important work related to the detection of hate speech, both by approaches based on computational methods and linguistic-based approaches.

Discursive practices with characteristics similar to hate speech are not new in Brazilian politics, mainly when the victim's gender differs from the normative standard. Recent cases presented do not represent the only ones that have occurred in our political scenario lately; however, it demonstrates a crescent wave of hate speech in this scenario, exposing the need for studies to combat such practices on different fronts. Here, our efforts are focused on cases of potential hate speech related to gender in the political sphere in virtual environments.

As exposed, in recent years, have shown a growing interest in applying natural language processing and natural language understanding to analyze and detect damaging speeches on the Internet, especially hate speech. Today's digital media ecosystem generates massive unstructured data streams, such as texts and documents available in various formats, thus posing a set of challenges related to their intuitive understanding. Although several attempts have been made to tackle the problem of detecting hate speech in social media by classifying texts written on it, several problems remain unsolved. Research development is still quite limited to the volume of resources available for each language. As we can see in the hate speech data monitoring[12], annotated datasets are scarce for most languages other than English.

Among the related works presented here, some different techniques and tools seek the processing and analysis of contents that improve hate speech detection in social networks based on distinct strategies. Between those strategies, current neural network methods have been shown to improve the results evaluated in separate models when applied to hate speech data. Nevertheless, in general, many of the works that aimed at detecting hate speech spent their efforts mainly on computational approaches (e.g., Badjatiya et al. (2017); Agrawal and Awekar (2018); and Arango, Pérez and Poblete (2019)), with linguistic heuristics in the background only. It is possible to observe that intolerant speeches present different nuances in their materialization. Thus, the exploration of textual data usually requires intensive linguistic analysis.

However, from machine learning-based approaches, the linguistic field is often

[12] See: <https://hatespeechdata.com>

considered only in terms of natural language processing tasks (e.g., tokenization, lemmatization, stemming, and part-of-speech tagging, among others), with few interdisciplinary means being applied. There are gaps linked to the complexities of the language studied and the limited use of the linguistic area when observing the hate speech identification literature from a computational point of view. The approaches are very much focused on lexical and syntactic issues when approaching them from a linguistic point of view, restricting research under fundamental semantic aspects; the procedure is always carried out in a non-automatic way, even in linguistic-based approaches, contributing to the need for very high wear of manual evaluators.

To the best of our knowledge, no computational approach explores the text classification process through frame semantic approaches within the hate speech spectrum. It does not even combine contributions made by the linguistic field, such as frame semantics. Moreover, the limitations are even more significant when considering Portuguese as a primary language. In this sense, we aim to provide a methodology based on computational methods to shed light on linguistic features inherent to intolerant speech.

4 A PROPOSAL FOR A COMPUTATIONAL LINGUISTIC APPROACH TO SUPPORT THE ANALYSIS OF THE DISCURSIVE CONFIGURATION OF VIOLENCE IN SOCIAL MEDIA

In this study design, we intend to present a formalization of linguistic heuristics proposed in the literature to support the automation of detecting hate speech in Brazilian Portuguese in social networks from the understanding given by Frame Semantics - considering the instantiation of symbolic violence frame. Therefore, the study covers the collection of data from social networks followed by a series of experiments and analysis of its contents. In a retrospective direction, it is intended to select by the presence of the outcome (hate speech), creating subgroups of data with the characteristics of this type of speech data with these absent characteristics. If there is a causal relationship between the exposure and the outcome, it is to be expected that the exposure is more often found in the group that presented the outcome. For this result to be valid, we have to consider certain assumptions, such as the adequate size of the sample studied, the minimization of bias occurrence (i.e., pre-inferences of the presence of harmful speech in a given text), statistical tests, etc. The sample selection exhibition will be based on ranking texts collected from social networks using the Perspective API (LEES et al., 2022). This API makes use of machine learning models to score the perceived impact that a comment could have on a conversation, also listing the probabilities that the analyzed content will be perceived as containing six different characteristics for the language Portuguese: toxicity, severe toxicity, identity attacks, insult, profanity, and threat.

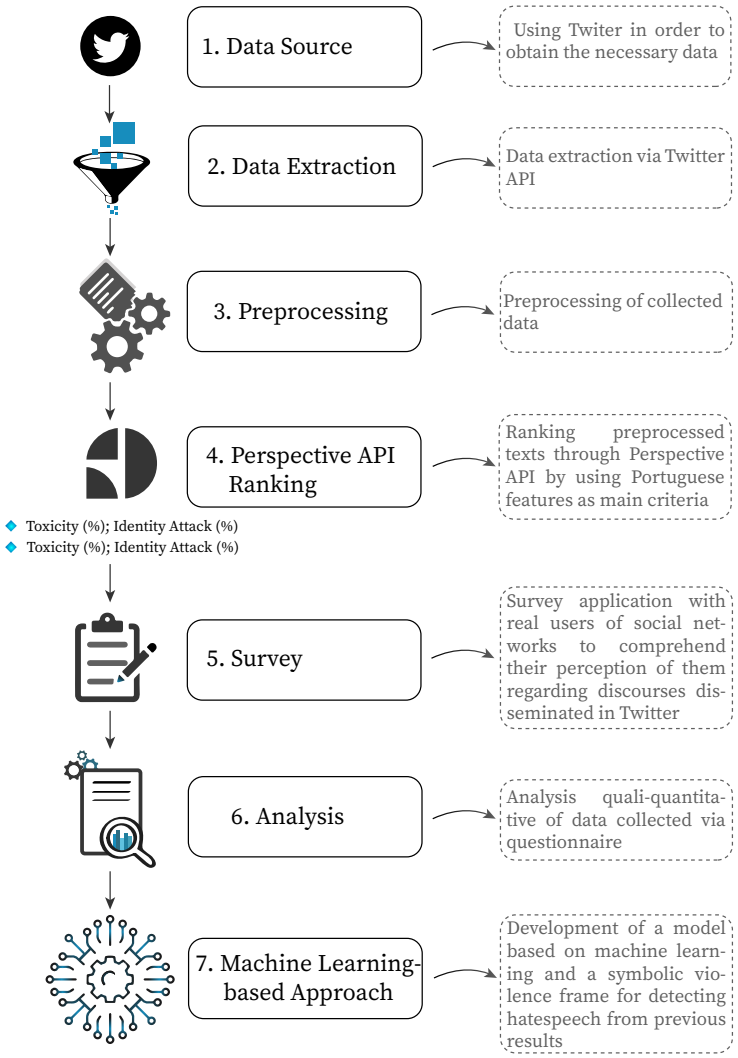
The core of this study is developed considering the following hypothesis:

> *The use of linguistic indicators that are characteristic of intolerant speeches (or even hate speech) linked to computational methods such as text classification can assist in the structuring of content disseminated in the virtual scope from the perspective of frame semantics, considering a frame of symbolic violence for that.*

This work has both a survey and an experiment with regard to its (WAZLAWICK, 2020) procedures. Following the research conditions previously mentioned: (1) proposal of a symbolic violence frame; (2) data collection (speeches) from social networks; (3) selection of data from the ranking provided by the use of the Perspective API; (4) assessment of the adequacy of the proposed frame for framing such speeches according to a formalization of linguistic heuristics proposed in the literature, used to automate the de-

tection of hate speech in the virtual environment, by analyzing the agreement between the users/evaluators collected through questionnaires. In Figure 4.1 a seven-stage workflow proposed to address such steps is presented.

Figure 4.1: Seven-stage workflow proposed



Source: The author (2023)

This workflow is composed by: (1) Data source selection; (2) Data extraction via API; (3) Preprocessing of collected data; (4) Ranking preprocessed texts through Perspective API by using Portuguese features as main criteria; (5) Survey application with real users of social networks to understand their perception of them regarding discourses disseminated in Twitter; (6) Analysis of data collected via questionnaire; (7) Development of a model based on machine learning and a symbolic violence frame for detecting hate speech from previous results.

4.1 Methodological proposal for data annotation and evaluation

Nowadays, social media and the amount of user-generated content continue to grow at a staggering rate (PÉREZ; GIUDICI; LUQUE, 2021). Microblogs have grown in popularity as a means of disseminating first-hand information (PIAO et al., 2020). Among the microblogging platforms, Twitter has become one of the most popular social networking platforms in recent years, providing a venue for millions of individuals to share their opinions. Twitter is one of the most standout social networking businesses, with one of the most influential platforms worldwide (AMNESTY INTERNATIONAL, 2021). Unlike other social networks, Twitter allows users to access, comment on, and contribute to all topics or threads (LI; SUN; DATTA, 2012). Thus, since users of these platforms can be seen as citizen journalists or sensor observations (NAGARAJAN et al., 2009), which is of great importance for the theme addressed in this work (political gender violence), and the ease of access to data, to conduct the present study we decided for the use of Twitter data.

In order to identify cases of intolerant discourse on social networks linked to political gender violence, tweets from women cis and transgender linked to Brazilian politics were extracted, as well as the replies received. Such contents date from the Brazilian electoral period of 2020, specifically from November 1st, 2020 - the beginning of the 2020 electoral campaign - to January 31, 2021 - the end of the first month of the term of the newly elected candidates. The data used in this work were extracted from Twitter[1] through its API (Application Programming Interface). In order to avoid violations of users' privacy, identity information was erased from the data. To standardize the data collected for further analysis, we applied some NLP preprocessing techniques such as lowercase texts, emoji, hyperlinks, hashtags and mentions remove, and the expansion of popular acronyms used on the Internet to their extended form. The goal of such a step was to keep only the texts from each tweet in a clean form.

To assess the adequacy of the proposed framework for framing discourses disseminated in social networks, according to a formalization of linguistic heuristics proposed in the literature used to automate the detection of hate speech in the virtual environment, a survey was proposed to be carried out through a questionnaire to real users of such

[1]Due to privacy issues and Twitter regulations, data collected for use in this study cannot be made publicly available. However, interested parties may contact the author directly to request access. Such requests will be evaluated, on a case-by-case basis, considering the twitter rules.

networks. To instruct the questionnaire respondents[2] we use the definition provided by Fortuna and Nunes (2018) as a definition of hate speech. Moreover, as inherent characteristics of hate speech to be considered, we made use of three characteristics (*Opposition: Us vs. Them (Themes and Opposition Figures)*, *Sanction for those who fail to comply with social contracts*, and *Passionate hatred and aversion to the different*) pointed out by Barros (2014), as well as a fourth characteristic empirically identified by us as potentially representative of intolerant discourses (*Fallacy intended to propagate hate*). The analysis of the agreement between users/evaluators regarding the classification of speeches in a spectrum of speech severity or their non-classification allows the perception of real users of social networks. It allows us to understand social network users' perception regarding speeches disseminated in such media.

A quali-quantitative approach was proposed to carry out the analysis of the collected data. By the qualitative aspect in the first moment, content analysis was performed using computational approaches (use of the Perspective API for data selection as described) that serve as a basis for a qualitative analysis of the observations and the questions opened in a questionnaire. On the other hand, the quantitative approach uses statistical analyses, such as frequency distributions, correlations, graphical representations, measures of dispersion, and measures of central tendency, thus seeking to observe how the present proposal fits or not with the expected outcomes. Such a proposal was carried out in an exploratory (i.e., aiming at building hypotheses) and diagnosis analysis (i.e., aiming to understand the causes of an event - framing data under the frame of symbolic violence - if possible answering surveys such as how and why). The methodology followed to analyze the social network data is described in Fragoso, Amaral and Recuero (2011) for studying social networks. Based on the premises of 'Social Networks Analysis' (SNA), two main steps are determined: the delimitation of the object and the data. Within the delimitation of the object, an attempt is made to trace the determination of a social network based on the researcher's object, that is, determining which elements the study seeks to be analyze in depth.In this step, the aim is to extract data from Twitter as already defined. Encoding information related to the user's identity will be applied to extracted tweets identifications data. Only the researcher members had access to this identification key to avoid violating participants' privacy.

When dealing with the data to be observed, it is proposed to rank texts collected using the Perspective API. This API makes use of machine learning models to score the

[2]See Appendix C.

perceived impact that a comment may have on a conversation, also listing the probabilities that the analyzed content will be perceived as containing six different characteristics, for Portuguese: toxicity, severe toxicity, identity attacks, insult, profanity, and threat. The Perspective API is widely used for toxicity evaluation and aims to establish safer online communication spaces (JIAWEN et al., 2022).

In order to study the use of linguistic heuristics to assist in the automation of hate speech detection in the virtual environment, a questionnaire[3] (a sample of this questionnaire can be found in Appendix E) was prepared to contain ten documents ranked through the attributes probabilities identified by the Perspective API, mainly the identity attack feature. Then, the evaluators, considering guidelines previously presented, carry out measurements through questions constructed using a rating scale with verbal descriptions that include extremes such as "*Totally fits*" to "*Does not fit*", and also the possibility to point out the absence of hate speech. This form of evaluation comes from the understanding that hate speech has several degrees, as put by Baider (2020). Also, using the same form, the objective was to evaluate the symbolic violence frame proposed through questions corresponding to the feasibility of adapting the frame as a way to describe texts such as those presented to the evaluator. Thus, speeches will then be checked in relation to their adequacy to a proposed frame (frame of symbolic violence further described in Section 4.2) in order to portray them according to a formalization of linguistic heuristics proposed in the literature (further discussed in Section 2.3), used to automate the detection of hate speech in the virtual environment. In all questions presented in the form, the evaluator will be able to leave comments that complement the answer provided. All evaluators considered for the research will be presented with a free and informed consent form (see Appendix D).

The dissemination of questionnaire proposed for the annotation of the dataset used in this work was released through social media (Facebook, Instagram, WhatsApp, and email lists), thus seeking to understand the perception of real users regarding discourses disseminated in such media. With the application of the questionnaire, through the analysis process previously described, it was expected to interpret content disseminated on Twitter (that could possibly be generalized to other similar social networks) considering a frame of symbolic violence. This understanding is accomplished by capturing the perception of evaluators to fine-tune the framing of such contents under this perspective, i.e., frame semantics. Analyzing the results of the classification given by users to validate the

[3]The questionnaire can be found in <https://survey-discurso-intolerante.formr.org>

proposed frame aims to understand a frame of symbolic violence, i.e., to analyze texts extracted from social networks considering the schematization of conceptual structures, beliefs, institutional practices that emerge from the daily experience, resulting in the representation of a situation, in this case, captured by the frame of symbolic violence by real users.

From there, our objective was to create a classification model that may serve as a computational linguistic-based approach to support the analysis of the discursive configuration of intolerance on social media (specially on twitter and other social medias with features similar to Twitter). In summary, our primary goal is: to evaluate the use of linguistic indicators associated with hate speech associated with computational methods such as the classification of texts by trained ML algorithms. For this purpose, as a way of enabling the framing of content disseminated on social networks considering a conceptual level representation model (frames), considering a symbolic violence frame as a way of proposing a means of interposing intolerant discourses. Thus, we intend to understand real users' perceptions when the speeches are disseminated in such media through an evaluation questionnaire to be presented to such users.

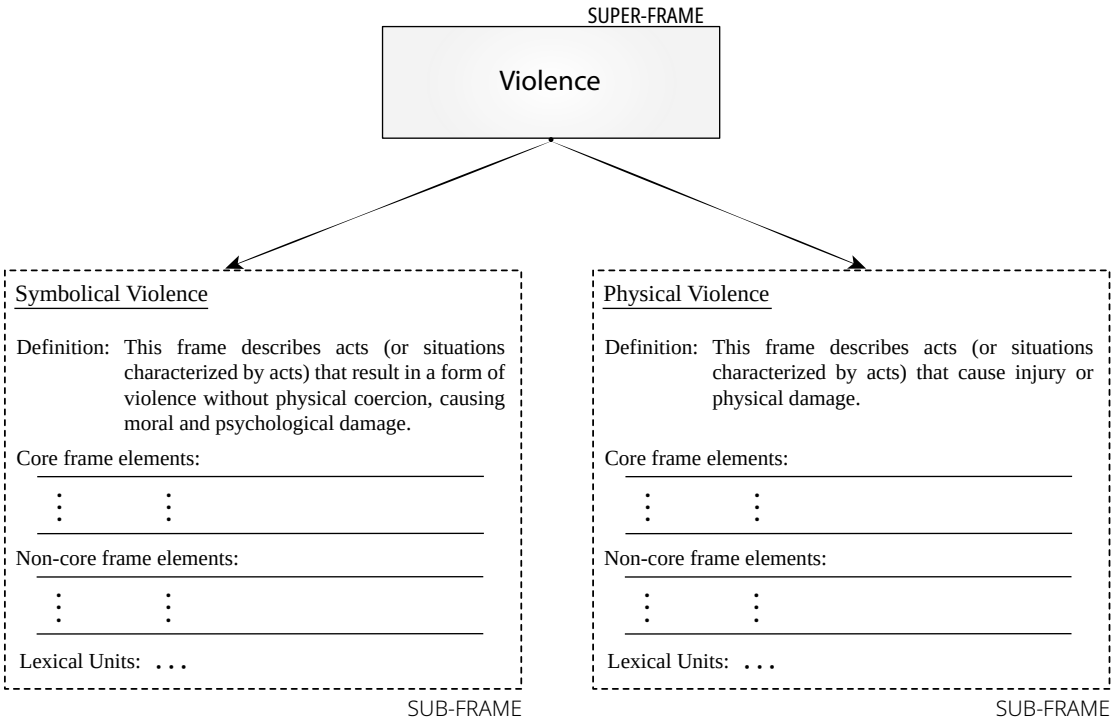
4.2 Symbolic Violence Frame

After observing studies aimed at the automatic detection of hate speech from different perspectives and approaches, it is possible to highlight that: (1) this is a study that requires the co-participation of different areas, not being restricted to just one research field; (2) there is a gap in the collaborative study between computation and linguistics, and from a computational point of view, linguistic features remain only in preprocessing steps, while from a linguistic point of view the automation of tasks is little explored. In summary, as stated when observing the hate speech detection literature from the computational point of view, there are gaps in the nuances of the analyzed language and the linguistic field's little use in the process. When considering from the linguistic point of view, the approaches are very much based on lexical and syntactic issues, limiting analyzes under extremely relevant semantic aspects; also, in linguistic-based approaches, the process is often carried out in a non-automatic way, leading to the high need of manual evaluators.

In this way, our proposal aims to contemplate a multidisciplinary approach by using a strategy based on frame semantics. In this work, we present a restructuring to the

current frame of *Violence* described in the FrameNet Brasil project. Here, we propose a structure where violence is seen as a super-frame that includes two sub-frames: a frame of physical violence and a frame of symbolic violence. Figure 4.2 presents an illustration of the new organization proposed, considering a new frame to represent symbolic violence. The definition of this frame follows the concept of symbolic violence by the sociologist Pierre Bordieu (see Section 2.2). This frame follows the structure used by FrameNet. The present frame proposal was originally designed considering a frame in Portuguese. However, for uniformity with the other contents described, we present its description/structure using English as the language in this section. A Brazilian Portuguese version can be found in Appendix B.

Figure 4.2: Restructuring proposed to the current violence frame. In this new organization, there is a super-frame called "Violence" from which two other sub-frames derive: "Symbolical Violence" (left side) and "Physical Violence" (right side).



Source: The author (2023).

Next, we present the proposed structure for such a frame:

DEFINITION

- This frame describes acts (or situations characterized by acts) resulting in a form of violence without physical coercion, i.e., symbolical violence, causing moral and psychological damage. The acts may involve an Aggressor injuring a Victim, or Aggressors causing harm to each other.

- They committed several defamations regarding the victim of the case.
- The prejudiced manifestations made in his speech made last week had great repercussions.

CORE FRAME ELEMENTS

- Aggressor the Aggressor is the person who causes harm to the victim
  - Your threats to his political competitors seem to know no limits.
- Aggressors Aggressors are a group of people who commit acts of violence jointly;
  - Attacks coordinated by institutions on virtual media have become a strategy to demoralize people and/or entities.
- Victim the Victim is being or entity that suffers the damage;
  - The increase in violence by virtual means against opponents of the government's proposal attracts international attention.

NON-CORE FRAME ELEMENTS

- Circumstances Circumstances describes a situation (at a particular time and place) that is specifically independent of the violent act or any of its participants;
- Interpretation Action or characteristic attributed to the victim whose aggressor addresses in a symbolically violent act (e.g., through attitudes and/or speeches referring to the victim);
- Containing_event Identifies the event in which the damage was caused;
- Purpose Identifies the Purpose for which the action that causes the damage is performed;
  - semantic_type: @state_of_affairs
- Frequency How often does violence occur;
  - Opposition figures are daily the target of violent acts on their social networks where they seek to discredit their image
- Degree Indicates the degree of violence committed

- semantic_type: @degree
- Election campaigns have been the stage for extreme spread of hatred.

- Iterations Iterations refer to the number of times that the violent act occurs;
- Means Means used by the aggressor to target a victim.
  - Massive messages of intimidation were received on their social media after their last speech.
- Manner The way the aggressor acts on the victim
  - semantic_type: @manner
- Time Identifies the Time in which the harmful event occurs.
  - semantic_type: @time

FRAME-FRAME RELATIONS:

- Inherits from: –
- Is Inherited by: –
- Perspective on: –
- Is Perspectivized in: –
- Uses: Cause_harm
- Is Used by: –
- Subframe of: Violence
- Has Subframe(s): –
- Precedes: –
- Is Preceded by: –
- Is Inchoative of: –
- Is Causative of: –
- See also: –

5 EXPERIMENTS

This chapter presents a full description of the experiments performed[1] on the proposed approach. All the stages required for the designed experiments are covered in depth, presenting, detailing, and discussing the main topics of this research.

5.1 Experiment Settings

We constructed our dataset to evaluate the proposed approach since no other datasets focused on gender violence or with correlation were found in Portuguese. Following the methodology described in Fragoso, Amaral and Recuero (2011) for SNA, in the first step, we set Twitter as the leading social network from where we seek to analyze in-depth. Nonetheless, it is important to state that violence and abuse against women are not restricted to any social media platform. However, research carried out by an observatory of political and electoral violence against candidates on social networks (REVISTA AZMINA; INTERNETLAB, 2021) showed that on Twitter, offensive comments and attacks on female candidates were more visible due to its open architecture.

The second step of the SNA deals with the data themselves, their collection, and intended analysis methods. As specified by Fragoso, Amaral and Recuero (2011), this depends on the analysis window intended to be performed. It is up to the researcher to select the moment and the variables that will be analyzed, which must be selected according to the problem that will be focused on. Thus, in defining our object of study, we chose to use data from Twitter, focusing only on texts disseminated through this platform aimed at women linked to politics. Images were disregarded in this study, even those containing texts.

5.1.1 Data Acquisition and Annotation

Creating a hate speech dataset typically entails annotating short documents such as web text, with or without providing context for the hate expressions (YANG; JANG; CHO, 2022). As it is a dense topic, data annotation was planned through questionnaires, presenting up to 10 tweets for annotation for each evaluator at a time. The question-

[1] All the codes developed in this work are available at <https://github.com/brendasalenave/phd_thesis>

naire was built using FormR (ARSLAN; WALTHER; TATA, 2019) as a support tool. This questionnaire was publicly disseminated through e-mail lists and social networks in Brazil. Considering the volume of collected data and forecasting the inherent difficulty of attracting volunteers for the annotation process, it was necessary to establish selection criteria for the data to be included in the annotation questionnaire. We ranked the collected texts using the Perspective API. After analyzing the definitions pointed out by the API, we considered that the identity attack metric was the closest to what we sought to analyze. So we established a threshold for this feature to list data to be annotated. The threshold set is 0.65 for the identity attack attribute. This value was chosen empirically, after testing different ones, aiming to select a possible amount of data to be annotated from a questionnaire in a non-exhaustive way and to cover data of potential interest. With the identity attack variable equal to and greater than 0.65, the selected dataset was composed of 120 instances, instances that met the defined inclusion criteria. Our goal was to evaluate each text by at least three different evaluators so that in case of disagreement between the classification of a tweet as intolerant speech or not, there would be yet another evaluation that would allow analysis by the majority.

In these guidelines for the annotation, we used the definition provided by Fortuna and Nunes (2018) as a definition of hate speech. Moreover, as inherent characteristics of hate speech to be considered, we made use of three characteristics (*Opposition: Us vs. Them (Themes and Opposition Figures)*, *Sanction for those who fail to comply with social contracts*, and *Passionate hatred and aversion to the different*) pointed out by Barros (2014), as well as a fourth characteristic empirically identified by us as potentially representative of intolerant discourses (*Fallacy intended to propagate hate*).

After being introduced to the guidelines, the data annotation process began. The annotator was then presented with a tweet written by a woman in politics and a reply sent to her referring to that tweet. The annotator was then asked to evaluate such a **reply** considering the previous guidelines presented. The first question related to the reply instance asked if the text could or could not be considered hate speech; the rater then evaluated through a rating scale ranging from 0 (no presence of hate speech) to 5 (high incidence of hate speech). If the evaluator identified the presence of hate speech in the analyzed content, i.e., with a response greater than or equal to two, the rater was then asked to indicate which characteristics were present in the text, which could be one or more. Raters could indicate other characteristics freely if necessary. In addition, the annotator was asked to assess the adequacy of a definition of a symbolic violence frame to tag the symbolic vi-

olence contained in the evaluated text. Such a definition is presented in the *c* item. The adequacy of this definition was also assessed using a five-point rating scale ranging from 0 (Not adequate) to 5 (Completely adequate). The reply classification process was repeated ten times per rater, presenting a different instance to be annotated.

Thus, we intended to follow the descriptive paradigm described by Röttger et al. (2021). As described by the authors, this descriptive paradigm supports annotator subjectivity, resulting in datasets that are granular surveys of individual beliefs. Thus, descriptive data annotation enables the capture and modeling of many beliefs.

5.2 Annotated Data Description

This section details and discusses the main outputs of the data annotation process, starting with the description of the profile of the respondents' participants in section 5.2.1, then following with a description of the intolerant speech characteristics' rating in section 5.2.2; next moving to a presentation of the Symbolical Violence Frame Adequacy in section 5.2.3; closing it with the presentation and discussion of the agreement analysis between annotators. Only data from users who answered the complete questionnaire were considered; that is, 83 answers (this corresponds to 74,77% of the total number of people who started to respond to the survey).

5.2.1 Respondents' Profile

As mentioned, in the first moment of the annotation questionnaire, the respondents were invited to answer a few questions indicating the social groups it fits. This step was carried out to outline the profile of the interviewees/respondents.

5.2.1.1 Gender

Respondents were asked which gender they identified with. The list of possible answers included the following options: "Female", "Male", "Other", and the option "I prefer not to disclose".

Figure 5.1 shows the distribution by declared gender of the questionnaire's respondents. Of the total of 83 respondents, 57.83% identified themselves with the male gender (labeled as "*Masculino*" in the figure), 38.55% identified with the female gender (labeled

as "*Feminino*" in the figure), 2.41% did not identify with any of the listed gender options (labeled as "*Other*" in the figure), and 1.20% preferred not to inform (labeled as "*Não*" in the figure).

Figure 5.1: Gender distribution of questionnaire respondents.

5.2.1.2 Race

Respondents were asked which race/ethnicity they identified with. The list of possible answers included the following options: "White" (a.k.a., caucasian), "Black", "Brown" (fruits of the miscegenation of ethnicities), "Yellow" (Asian descent), "Indigenous", and the option "I prefer not to disclose". The options presented were extracted from the Brazilian Institute of Geography and Statistics (*Instituto Brasileiro de Geografia e Estatística* – IBGE).

Figure 5.2 shows the distribution by age group of respondents to the questionnaire. Of the 83 respondents, 65.06% declared themselves as white (labeled as "*Branca*" in the figure), 15.66% declared themselves as black (labeled as "*Preta*" in the figure), 15.66% declared themselves as brown (labeled as "*Parda*" in the figure), 2.41 declared themselves as yellow (labeled as "*Amarela*" in the figure), and 1.20% preferred not to inform (labeled as "*Não*" in the figure).

5.2.1.3 Education Level

Respondents were asked about their highest level of education. The list of possible answers included the following options: "Basic Education", "High School", "Technical Education", "Higher Education", "Specialization", "Master's", and "Doctorate".

Figure 5.2: Race distribution of questionnaire respondents

Figure 5.3 shows the distribution by age group of respondents to the questionnaire. Of the total of 83 respondents, 40.96% had higher education (labeled as "*Superior*, 22.89% had a doctorate (labeled as "*Doutorado*" in the figure), 22.89% had a master's degree (labeled as "*Mestrado*" in the figure), 10.84% secondary education (labeled as "*Medio*" in the figure), and 02.41% specialization (labeled as "*Especialização*" in the figure).

Figure 5.3: Education level distribution of questionnaire respondents.

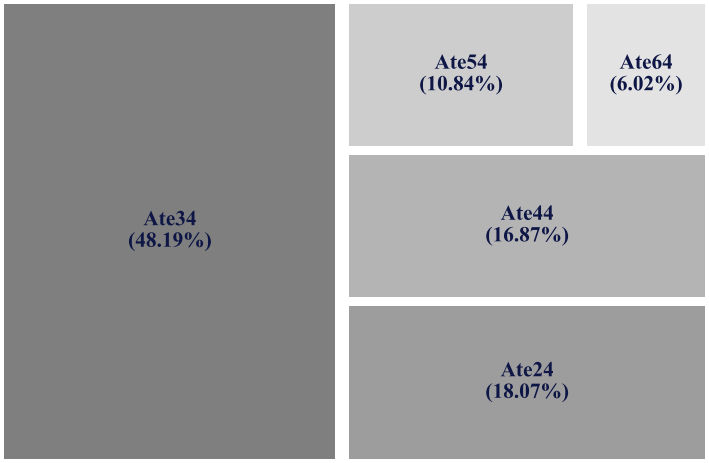
5.2.1.4 Age Group

Respondents were asked which age group they fell into. The list of possible answers included the following age groups: "From 18 to 24 years old", "From 25 to 34 years old", "From 35 to 44 years old", "From 45 to 54 years old", "From 55 to 64 years old",

"65 years or older", and the option "I prefer not to disclose".

Figure 5.4 shows the distribution by age group of respondents to the questionnaire. Of the total of 83 respondents, 48.19% were in the age group between 25 and 34 years old (labeled as *Ate34* in the figure), 18.07% in the age group between 18 and 24 years old (labeled as *Ate24* in the figure), 16.87% in the age group between 35 and 44 years old (labeled as *Ate44* in the figure), 10.84% in the age group between 45 and 54 years old (labeled as *Ate54* in the figure), and 6.02% in the age group over 65 years old (labeled as *Ate64* in the figure).

Figure 5.4: Distribution between Age Group of questionnaire respondents

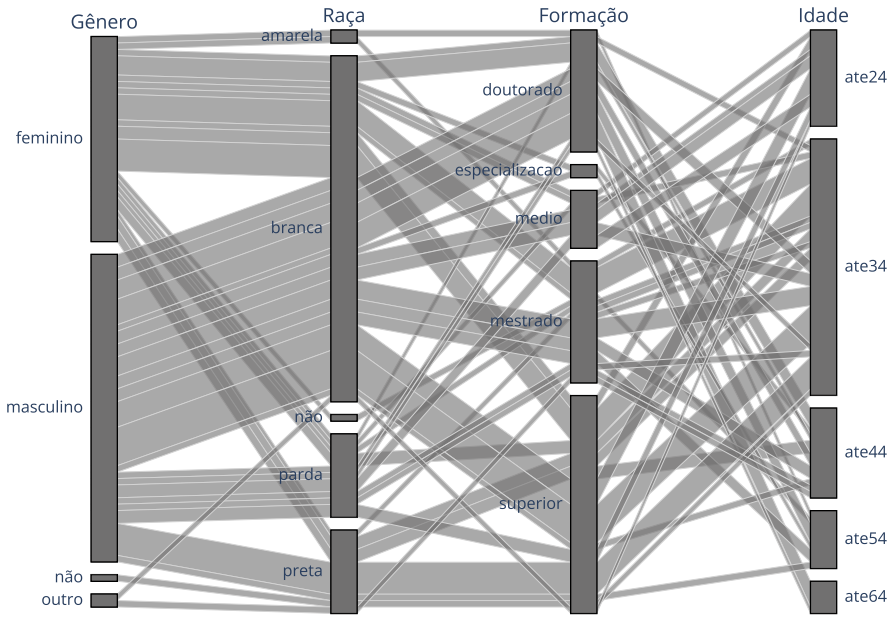


5.2.1.5 Crossing respondents profile features

Figure 5.5 presents through a parallel categories chart the multi-dimensional categorical relationships between the respondents' features collected in the questionnaire. Each variable in the data set is represented by a column of rectangles, where each rectangle corresponds to a discrete value taken on by that variable. The relative heights of the rectangles reflect the relative frequency of occurrence of the corresponding value. Here, each vertical bar shows the overall winning percentages, and following thicker lines reveals where strong co-occurrences lie. Combinations of category rectangles across dimensions are connected by ribbons, where the height of the ribbon corresponds to the relative frequency of occurrence of the combination of categories in the data set. This plot gives one an overview of the questionnaire respondents' profiles.

Aiming to go further in the analysis of multi-dimensional categorical relationships, Figure 5.6 presents a complement to the previous one, adding the average rate (la-

Figure 5.5: Multi-dimensional categorical relationships between the respondents features collected in the questionnaire



beled in the figure as "*Média*)" provided by the different groups of profile for each tweet reply labeled in the questionnaire. The rating of intolerant speech in each tweet reply was measured by applying a rating scale, where 0 indicated the absence of intolerant speech and 5 considered a high incidence (further discussed in Section 5.2.2). Through this plot, one can note that:

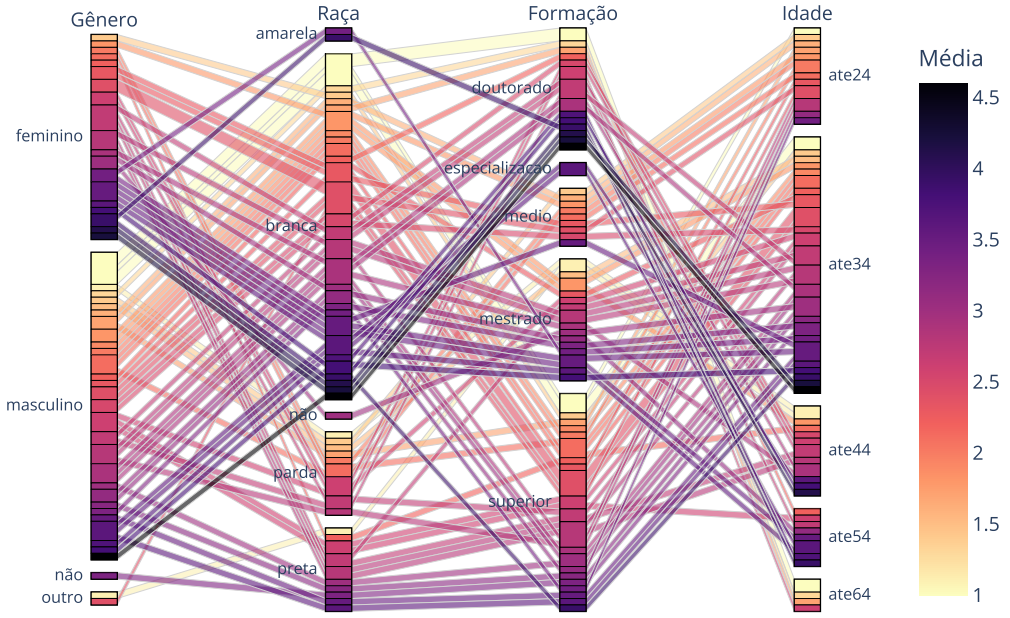
- People who identify with the female gender have a higher average in the evaluation of speeches that are intolerant than those who identify with other genders or do not prefer to declare; that is, they tend to evaluate speeches as being more intolerant.
- People who identify with the masculine gender have a lower evaluation average than those who identify with other genders or do not prefer to declare; that is, they tend to evaluate the speeches more leniently.
- Among people who identify as male, those who also identify as black have a higher rating average than those who identify with other races.
- People who identify as white, in general, tend to rate speeches more leniently than those who identify with other races or choose not to state.

- People who identify as black, in general, tend to evaluate speeches more strongly, with a higher average, than those who identify with other races or choose not to declare.

- People with a specialization, master's, or doctorate as a higher education level have a higher average rating; that is, they considered the speeches evaluated as more intolerant than those with secondary education or higher education as a higher current education degree.

As stated, the descriptive paradigm followed fosters annotator subjectivity, resulting in datasets that are granular surveys of individual beliefs (RÖTTGER et al., 2021). Thus, insights regarding the views of annotators or the wider community they may represent may be derived from the distribution of data labels across annotators and instances. Hence, based on what is observed here, it is worth emphasizing the impact of the individual background carried out by one at the moment of rating and labeling the data.

Figure 5.6: Multi-dimensional categorical relationships between the respondents features collected in the questionnaire considering average rates

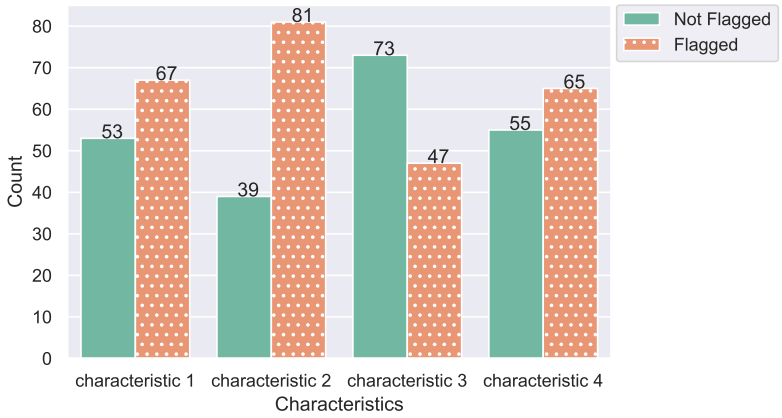


5.2.2 Intolerant Speech Characteristics'

For each annotated tweet, we analyzed whether each feature was flagged or not. The characteristics *Opposition: Us vs. Them (Themes and Opposition Figures)*, *Sanction for those who fail to comply with social contracts*, *Passionate hatred and aversion to the different*, and *Fallacy intended to propagate hate* were mapped to characteristic 1, characteristic 2, characteristic 3, and characteristic 4, respectively, for readability purposes. This notation will be maintained throughout this work. Figure 5.7 presents the presence count of each feature for the set of tweets. The characteristics considered flagged pointed out by at least one annotator.

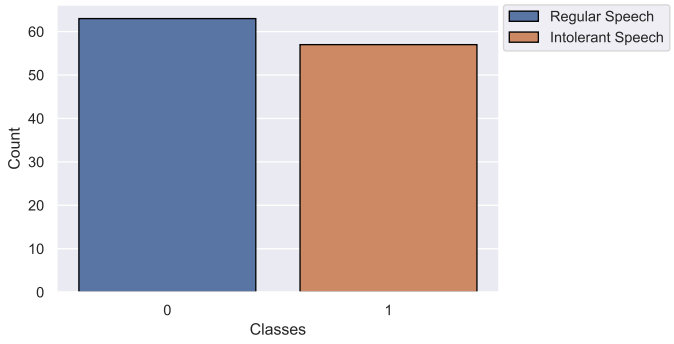
As seen in Figure 5.7, characteristic 4, i.e., the characteristic proposed by us, was pointed out by the questionnaire respondents in almost half of the entries. This data validates the use of this characteristic as a potential indicator of linguistic character to identify hate speech. The annotators suggested no extra features.

Figure 5.7: Characteristics presence count



Aiming to further analyze the role of the characteristics pointed out in the annotation of what is or is not intolerant speech, we seek to explore them in the context of predictive analysis to identify such instances. Considering that the measurement of the presence of intolerant speech was made using a rating scale, where 0 indicated the absence of intolerant speech and 5 considered a high incidence, we considered the average of the values pointed out by the annotators, and defined a threshold of 2.5 to indicate possible intolerant speeches. Thus, tweets with an average rating below 2.5 were considered non-intolerant, and tweets with an average above this threshold were considered intolerant. Figure 5.8 shows the class distribution of this new variable. As is possible to note, it leads to a low unbalance.

Figure 5.8: Intolerant speech variable distribution



5.2.3 Symbolical Violence Frame Adequacy

Our proposal aims to consider a strategy based on frame semantics. Studies in computational linguistics, especially focusing on computational semantics, design meaningful representations and establish strategies for automatically assigning and reasoning those representations (ERK, 2018). As stated by Ruas et al. (2020), the association of words in a sentence often tells us more about the underlying semantic content of the document than its literal words individually. In this sense, some methods build semantic representation based on the content analyzed, i.e., the way text components relate is observed as a pattern to formalize a model that provides a broader understanding of the content to humans.

One of the objectives of carrying out this questionnaire was to assess the adequacy of the proposal to define a frame of symbolic violence in the context of texts considered violent by the respondents. In order to verify this adequacy, a correlation study was carried out between the evaluation received by a tweet and the adequacy level of the proposed frame definition as a way of framing the respective tweet. Correlations were obtained using Pearson's Correlation as a basis for calculation. Thus, correlations equal to 1 were considered perfect correlations; between 0.9 and 1 were considered very strong; between 0.6 and 0.9 were considered strong; between 0.3 and 0.6 were considered moderate; between 0.1 and 0.3 were considered weak; and correlations between 0.0 and 0.1 were considered null. The exact correspondence occurred for negative values.

When observing the correlations obtained, it was noted that in most cases (72%), the correlations were strong, very strong, or even perfect. Thus, it is possible to infer that, in cases where the tweets were considered more violent, the definition proposed for the symbolic violence frame served to frame them correctly; as well as that the less intolerant the tweets were considered, the less the proposed definition for the symbolic

violence frame would fit them. In short, it is possible to assume that the current definition proposed for the symbolic violence frame is sufficient to identify more violent cases and fails to include cases of lower incidence of violence, thus needing to be revised for these cases.

5.2.4 Agreement Analysis

Annotations present subjective classifications; in this case, the topic addressed by itself can also be considered entirely subjective. In these cases, the study of Concordance Analysis between Evaluators by Attributes is used to evaluate the consistency and correction of subjective evaluations. Inter-annotator agreement (IAA) measures how well two (or more) annotators can make the same annotation decision for a specific category. It is a vital part of the validation and reproducibility of classification results. From the inter-annotator agreement measure, one derive two things: (1) How easy was it to define the category clearly: the annotator's criteria were quite explicit, implying that it is possible to provide a well-defined picture of the category for each type of item to the annotator; (2) How reliable is the annotation: when the inter-annotator agreement was low, it was difficult for the annotators to agree on which items belonged to a category and which did not. Such a category may be exciting from a qualitative standpoint, but it would be tough to include it in a quantitative assessment.

Unlike other inter-rater reliability metrics, this may handle various sample sizes, categories, and numbers of raters. Also, it applies to any measurement level (i.e., nominal, ordinal, interval, or ratio). For example, by measuring IAA through the Kripperdorff coefficient between the raters to the set of tweets and treating the data as ordinal, we achieved an alpha value of 0.429. According to Krippendorff (2004), an alpha in the range between 0.67 and 0.80 indicates low reliability. Ideally, it should be over 0.80. The alpha coefficient represents a deficient agreement between raters in our case. Considering the subjectivity of the theme, this sets up an apparent difficulty in objectively classifying the data. As exposed by Kocoń et al. (2021) in tasks involving subjectivity in annotation, the agreement rarely exceeds a moderate level without an experienced team of annotators in such tasks. And in situations where there is a high level of agreement, this is frequently due to the annotators' freedom of expression being limited.

Different from datasets which consist of clear hate and non-hate that can have very high levels of inter-annotator agreement, even with minimal guidelines (RÖTTGER

et al., 2021), in the annotation paradigm followed, like a fine-grained survey, it captures a variety of beliefs in data labels. The distribution of data labels between annotators and instances might thus reveal insights regarding annotators' ideas or the wider community they may represent (RÖTTGER et al., 2021). Despite the low agreement achieved in the annotation, we further investigated the characteristics pointed out by the annotators in the speeches and their relationship with the identification or not of intolerance.

5.3 Intolerant Speech Characteristics' Classification

Next, we present some experiments with classification attempts aiming to identify intolerant speech characteristics and where a speech might be intolerant or not. To this end we logistical regression as a baseline model. Logistic regression is a statistical model used in predictive analysis as a way to determine the probability of an event happening. It shows the relationship between resources and then calculates the probability of a given outcome (LAVALLEY, 2008). Table 5.1 presents the output of a Generalized Linear Model (GLM) Regression made on top of the dataset considering the four characteristics as the independent variable and the variable *intolerant* as the dependent one.

Through these results, we can observe that all the characteristics were statistically significant ($p < 0.05$) for the accomplishment of the prediction of the target attribute. Through the coefficients of each variable, it is possible to conclude that statistically, with a confidence interval of 95%: (1) the chance of a tweet that presents characteristic 1 being considered intolerant is 8.23 times greater than the chance of a tweet that does not have this feature; (2) the chance of a tweet that presents characteristic 2 being considered intolerant is 9.84 times greater than the chance of a tweet that does not present this characteristic; (3) the chance of a tweet that presents characteristic 3 being considered intolerant is 3.05 times greater than the chance of a tweet that does not present this characteristic; and (4) the chance that a tweet that has characteristic 4 will be considered intolerant is 3.05 times greater than the chance of a tweet that does not have this characteristic.

Considering the role played by each of the characteristics to classify a speech as potentially intolerant, we then explore text classification techniques, aiming to identify the presence of each characteristic. Thus, by adding one more level of classification, in addition to being able to identify hate speech, we hope to be able to predict with a certain reliability, the reasons (characteristics) that make up this classification. The following subsections detail the prediction of the manifestation of each characteristic, based on the

text present in each tweet.

5.3.1 Characteristic 1

We started setting up a baseline classification. As a baseline model, choose Logistic Regression for its versatility. Using this model, for classifying characteristic 1 presence from the tweet's text, we calculated the score metrics in five ways: Bag of words, TF-IDF with 1-gram, TF-IDF with 2-gram, Word2Vec CBOW, and Word2Vec Skip-Gram. Table 5.2 presents the results achieved for each approach. The best results for this characteristic were achieved using the Word2vec CBOW model for preprocessing. This approach achieved an F1-score of 0.635, a value to be used to compare with improved methods.

Table 5.1: Generalized Linear Model Regression Results

Dep. Variable:	['intolerant[no]', 'intolerant[yes]']	**No. Observations:**	120
Model:	GLM	**Df Residuals:**	115
Model Family:	Binomial	**Df Model:**	4
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-46.107
Date:	Wed, 11 May 2022	**Deviance:**	92.215
Time:	09:26:57	**Pearson chi2:**	110.
No. Iterations:	6	**Pseudo R-squ. (CS):**	0.4595
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	4.8709	0.994	4.898	0.000	2.922	6.820
characteristic_1[T.yes]	-2.1076	0.563	-3.742	0.000	-3.211	-1.004
characteristic_2[T.yes]	-2.2860	0.688	-3.325	0.001	-3.634	-0.938
characteristic_3[T.yes]	-1.1168	0.545	-2.048	0.041	-2.185	-0.048
characteristic_4[T.yes]	-2.2559	0.566	-3.989	0.000	-3.364	-1.147

Table 5.2: Characteristic 1 classification baseline

Preprocessing	**Precision**	**Recall**	**F1-score**	**Accuracy**
Bag of words	0.293	0.542	0.381	0.542
TF-IDF 1-gram	0.525	0.542	0.442	0.542
TF-IDF 2-grams	0.525	0.542	0.442	0.542
Word2vec CBOW	**0.709**	**0.667**	**0.635**	**0.667**
Word2vec Skip-Gram	0.623	0.625	0.623	0.625

Aiming to improve the achieved metrics, we performed a series of tests with varying combinations of preprocessing and models. The preprocessing methods tested were: count vectorize, TF-IDF with 1 and 2 grams, and Word2Vec CBOW and Word2Vec Skip-Gram. Given the popularity of these well-established approaches in the NLP literature, and following the principle of Occam's blade (WAZLAWICK, 2020), we chose to make use of these methods, thus seeking to use simpler approaches that meet our purpose in this work. For the construction of the classification models, the following were used: Bernoulli Naive Bayes (NB), Complement Naïve Bayes, Logistic Regression, LightGBM, and XGBClassifier. The same principle of simplicity together with consideration of the performance of algorithms in text classification tasks in general was considered in the selection process. Bearing in mind that there is an imbalance in the *characteristic 1* variable (67 instances where the characteristic was flagged and 53 instances where it was not), we still chose to test such combinations of preprocessing and models with the data rebalanced. To undersample our dataset, we used NearMiss algorithm (YEN; LEE, 2006); thus, from the initial set of 120 observations, 106 were kept. We used the SMOTE (Synthetic Minority Over-sampling Technique) algorithm to oversample the dataset. This approach resulted in a total of 134 instances in the dataset.

From those combinations, the best results were achieved when using TF-IDF 2-grams with the Bernoulli Naïve Bayes classifier in the oversampled dataset. From that we achieved 0.870 of precision; 0.741 of recall; 0.756 of F1-score; and 0.741 of Accuracy; a clear improvement from the baseline model results ($> 10\%$). The best results achieved for the unbalanced, undersampling balanced, and oversampling balanced data are presented in Table 5.3.

Initially, more robust algorithms were not taken into account because, due to the

Table 5.3: Best results for the classification of characteristic 1

Balancing	Preprocessing	Model	Precision	Recall	F1-score	Accuracy
Unbalanced	TF-IDF 1-gram	Bernoulli NB	0.962	0.583	0.699	0.583
Undersampling	Word2vec Skip-Gram	Bernoulli NB	0.901	0.636	0.692	0.636
Oversampling	**TF-IDF 2-grams**	**Logistic Regression**	**0.870**	**0.741**	**0.756**	**0.741**

dataset size, their application could more easily lead to overfitting and less interpretability of the results. For the classification process of the next three remaining characteristics, the test protocol followed in this section was maintained.

5.3.2 Characteristic 2

For characteristic 2, the baseline results achieved with each approach are presented in Table 5.4. The best results for this characteristic classification were achieved by using Word2vec Skip-Gram in the preprocessing step. This approach achieved an F1-score of 0.653, a value to be used to compare with improved methods.

Table 5.4: Characteristic 2 classification baseline

Preprocessing	Precision	Recall	F1-score	Accuracy
Bag of words	0.444	0.667	0.533	0.667
TF-IDF 1-gram	0.424	0.583	0.491	0.583
TF-IDF 2-grams	0.435	0.625	0.513	0.625
Word2vec CBOW	0.633	0.667	0.630	0.667
Word2vec Skip-Gram	**0.648**	**0.667**	**0.653**	**0.667**

Undersampling our dataset to classify characteristic 2, 78 instances were kept from the initial set of 120 observations. By oversampling it, 162 instances were set to use. The best results achieved for the unbalanced, undersampling, and oversampling balanced data are presented in Table 5.5. Compared to the baseline results, oversampling our dataset results in an even more significant improvement than the other approaches (unbalanced and the undersampling data).

Table 5.5: Best results for the classification of characteristic 2

Balancing	Preprocessing	Model	Precision	Recall	F1-score	Accuracy
Unbalanced	TF-IDF 1-grams	Bernoulli NB	0.964	0.708	0.796	0.708
Undersampling	Word2vec CBOW	Logistic Regression	0.758	0.688	0.699	0.688
Oversampling	**TF-IDF 2-grams**	**Logistic Regression**	**0.925**	**0.909**	**0.910**	**0.909**

5.3.3 Characteristic 3

For the characteristic 3, the baseline results achieved with each approach are presented in Table 5.6. The best results for this characteristic classification were achieved using by TF-IDF with 1-gram in the preprocessing step. This approached achieved an F1-score of 0.605, a value to be used in comparison with improved methods.

Table 5.6: Characteristic 3 classification baseline

Preprocessing	Precision	Recall	F1-score	Accuracy
Bag of words	0.512	0.583	0.514	0.583
TF-IDF 1-gram	**0.604**	**0.625**	**0.605**	**0.625**
TF-IDF 2-grams	0.391	0.625	0.481	0.625
Word2vec CBOW	0.565	0.583	0.570	0.583
Word2vec Skip-Gram	0.565	0.583	0.570	0.583

Undersampling our dataset to classify characteristic 3, 94 instances were kept from the initial set of 120 observations. By oversampling it, 146 instances were set to use. The best results achieved for the unbalanced, undersampling, and oversampling balanced data are presented in Table 5.7. Compared to the baseline results, oversampling our dataset results in an even more significant improvement than the other approaches (unbalanced and the undersampling data).

Table 5.7: Best results for the classification of characteristic 3

Balancing	Preprocessing	Model	Precision	Recall	F1-score	Accuracy
Unbalanced	TF-IDF 2-grams	Logistic Regression	1.000	0.625	0.769	0.625
Undersampling	Count Vectorize	LightGBM	1.000	0.526	0.690	0.526
Oversampling	**TF-IDF 2-grams**	**Bernoulli NB**	**0.902**	**0.867**	**0.869**	**0.867**

5.3.4 Characteristic 4

For characteristic 4, the baseline results achieved with each approach are presented in Table 5.8. The best results for this characteristic classification were achieved using bag of words to text representation in the preprocessing step. This approach achieved an F1-score of 0.615, a value to be used to compare with improved methods.

Table 5.8: Characteristic 4 classification baseline

Preprocessing	Precision	Recall	F1-score	Accuracy
Bag of words	**0.625**	**0.625**	**0.615**	**0.625**
TF-IDF 1-gram	0.534	0.542	0.529	0.542
TF-IDF 2-grams	0.581	0.583	0.565	0.583
Word2vec CBOW	0.515	0.500	0.493	0.500
Word2vec Skip-Gram	0.493	0.500	0.493	0.500

Undersampling our dataset to classify characteristic 4, 110 instances were kept from the initial set of 120 observations. By oversampling it, 130 instances were set to use. The best results achieved for the unbalanced, undersampling, and oversampling balanced data are presented in Table 5.9. Compared to the baseline results, oversampling our dataset results in an even more significant improvement than the other approaches (unbalanced and the undersampling data).

Table 5.9: Best results for the classification of characteristic 4

Balancing	Preprocessing	Model	Precision	Recall	F1-score	Accuracy
Unbalanced	Count Vectorize	Bernoulli NB	0.962	0.583	0.699	0.583
Undersampling	TF-IDF 2-grams	Logistic Regression Bayes	0.686	0.682	0.682	0.682
Oversampling	**TF-IDF 1-grams**	**Bernoulli NB**	**0.876**	**0.769**	**0.782**	**0.769**

5.4 Intolerant Speech Classification and Validation

Next, after having trained classification models to identify each of the characteristics of intolerant speech, we moved on to a classification model focused on identifying potential intolerant speeches based on the presence or absence of the investigated characteristics. In this way, we elaborate a multi-layer classifier where the output of the classification of the best model applied for identifying the characteristics provides the input of a general classifier. This approach was followed given the importance of the role of each characteristic in classifying a speech as potentially intolerant (further discussed in Section 5.3). To this end, we selected a different data set, collected in the same period. Those tweets were also addressed to the same women with a value of identity attack identified by Perspective API over the 0.65 threshold. The classification algorithm did not previously know all the tweets analyzed in this step.

Figure 5.9 presents the presence count of each characteristic pointed out by the classification for this new set of tweets. As can be noted, different from the previous case, annotated by humans, there was a great imbalance in the identification of characteristics in this classification. While for characteristics 1 and 4 there was a significant presence pointed by the algorithm, for characteristic 3 few were the cases that were identified as presenting such characteristics.

After identifying the possible characteristics that are present in each tweet, we moved on to a more general classification. As done in the previous configuration, in this second stage, tweets were classified as regular or potentially intolerant based on the characteristics identified for each one. Figure 5.10 shows the class distribution predicted by this step.

We then applied a second questionnaire to evaluate the output of such a classification. At this stage, we invited people who participated by answering the first questionnaire

Figure 5.9: Characteristics presence count for unseen data

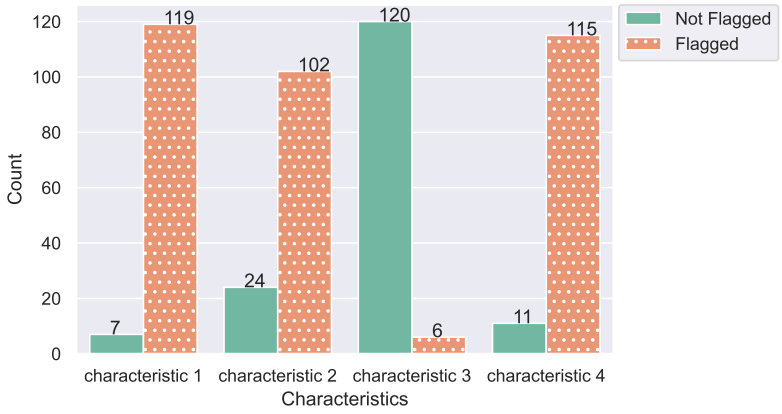
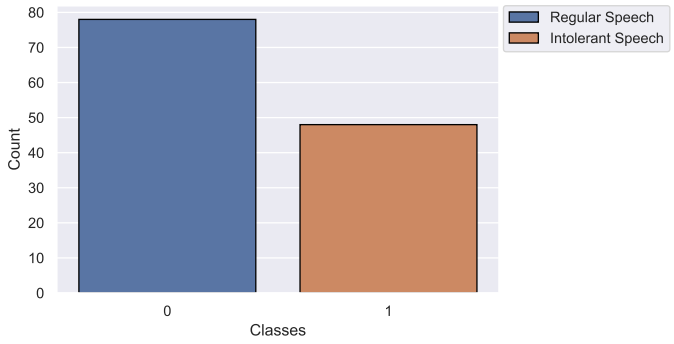


Figure 5.10: Intolerant speech variable distribution for unseen data

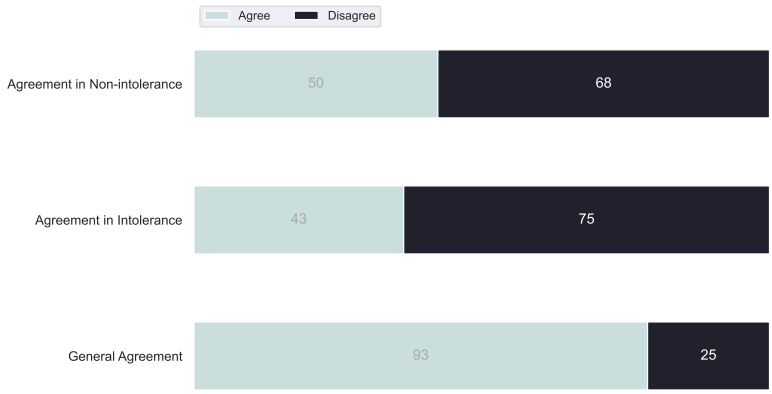


to collaborate. In this questionnaire, after the characteristics of intolerant speech considered in this study were again presented, the evaluators were invited to evaluate texts that were automatically annotated and indicate whether the classification and the characteristics pointed out by it are adequate (or not). This validation process was developed over three days through a sequential form prepared with the support of FormR (ARSLAN; WALTHER; TATA, 2019). It was decided to split the questionnaire in this way because the volume to be evaluated was too large for it to be done by one person in a single day, and it would require a high amount of time (more than one hour) of concentration and dedication to the process. Thus, on each day, the respondent would receive a set of tweets and an automatically generated annotation for each one of them and was asked about the correctness of the classification presented when indicating the presence/absence of hate speech and even if the characteristics of this type of speech indicated and that led to this decision are coherent. At the end of the set presented each day, there was also a question related to the adequacy of the frame of symbolic violence proposed by us for the contemplation of speeches classified as intolerant in the evaluated set. A subset containing 13 tweets was removed from this validation stage, as few tweets from one of the profiles considered in this study met the inclusion criteria; the tweets included in the previous

stage of the experiment ended up being repeated and could add bias to the evaluation.

From the 118 classified and evaluated instances results, at least two evaluators agreed that there was no intolerance in 50 instances, and at least two evaluators agreed that there was intolerance in 43 instances. In 93 of the 118 instances, at least two evaluators agreed regarding the presence or absence of intolerance. However, in only 64 cases (54.24%), the classification generated by the algorithm matched what at least 2 of the evaluators pointed out. This agreement in classifications is summarized in Figured 5.11.

Figure 5.11: Classification agreement between annotators in validation.



In this experiment, in addition to the general classification of a tweet as being intolerant or not, the respondent was also presented with the characteristics considered inherent to this type of speech, which influenced the classification decision-making. Thus, the respondents of the validation questionnaire were also invited to assess whether the characteristics suggested by the automatic classification matched the characteristics identified in the evaluated text. Figure 5.12 shows the general agreement of each user when evaluating the suggested characteristics. It is possible to notice that user 1 disagrees with the characteristics suggested by the classification in most cases. In contrast, user 2 tends to agree most of the time with the suggested classification. User 3, on the other hand, has a slight tendency to disagree with the suggested characteristics, but in several cases, this user agrees.

Thus, we also sought to analyze the number of times a characteristic was suggested by the classification algorithm and the number of times in which, in case of disagreement with what was proposed, the annotator user manifested to identify a particular characteristic in the analyzed text. Figure 5.13 presents the count of instances identified by the automatic classification and identified by the respondents for each characteristic.

The validation experiment results do not seem to be prospective at first sight. However, it should be noted that attempts based on manually labeled data carry bias. The

Figure 5.12: Characteristics classification agreement between annotators in validation.

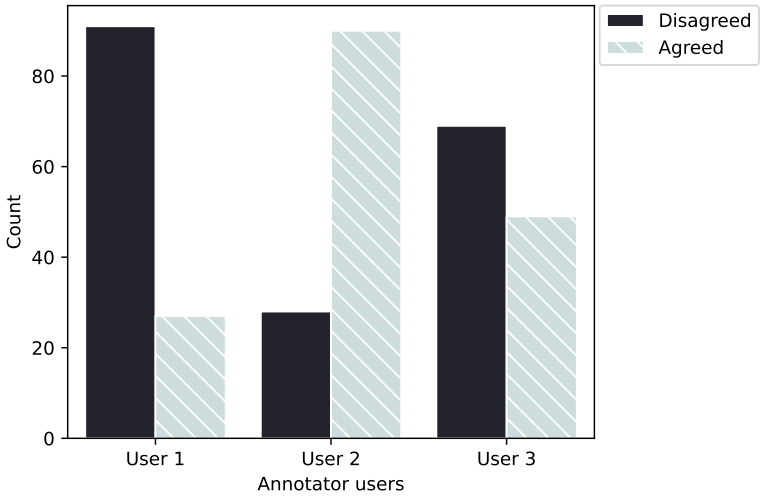
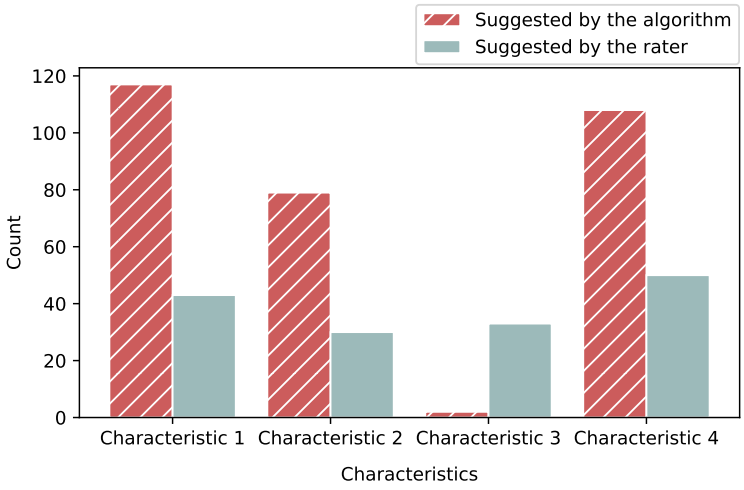


Figure 5.13: Occurrence count of identified and suggested characteristics.



same occurs to those focused on assisting in a more rigorous defense strategy to deal with possibly problematic inputs and contribute to a more secure deployment environment for language models because different annotators will have different perspectives and interpretations of the same data. Thus, this bias can be seen as a load deposited by its annotators from their experiences. That is, there will always be a portion of data bias, as these reflect the population's perception that contributed to the annotation, influencing the subsequent use of these data.

5.5 Discussion, challenges, and limitations

The goal of building and making available a dataset related to gender political violence that would assist researchers interested in classifying this type of content (both computationally and through linguistic features) has been achieved. As far as we know, this dataset is unprecedented in Brazil and can potentially be used as a tool and benchmark studies, such as Machine Learning algorithms that can be used to predict where a text content might contain intolerant speech related to political gender violence. The difficulties in building this data set were mainly related to instructing the annotators properly, namely, the definitions adopted to classify and apply the characteristics related to intolerant speech. Hate speech covers a variety of topics, targets, and situations. The dataset described in this thesis was built to focus on detecting hate speech related to political gender violence. Thus, its use must consider the application context. Although this topic may cover many others characterized as intolerant discourse (racism, misogyny, religious intolerance), generalizing the use of data to detect intolerant discourse in different contexts was not foreseen.

Although the proposal of this dataset is of great value for the training of algorithms to identify intolerant discourses linked to gender political violence, there are still several challenges to be faced. One example is the presence of ideological biases. Given the context that permeates the data and the period they were annotated (first quarter of 2022 - election year), we emphasize the possibility that such data may reflect annotators' ideological biases, even with the annotation protocol followed. However, as these are observations related to language, we must remember that language is not neutral. Any view of language is ideological because it reflects a specific perspective and emerges within a particular context (ROSA; BURDICK, 2016).

We highlight the contributions of this dataset from three fronts: linguistic, compu-

tational, and social fields. This is the first annotated Portuguese datase related to gender political violence instances. Also, it is the first one to consider linguistic characteristics associated with intolerant/hate speech. This dataset contributes to validating the use of a fourth feature that can be associated with such discussions (and possibly others): the use of fallacies. In terms of computational contributions, preliminary experiments have shown that, despite being a small dataset, it can support identifying hate speech instances and predict cases where each characteristic may or may not be found to support the decision-making in the final classification as hate speech. Finally, combating gender-based political violence is vital for maintaining democracy. Thus, approaches that help to identify potentially intolerant instances linked to the topic are of outstanding social contribution. We hope this dataset to become helpful to the computational linguistics community and may serve as a starting point for future research in this field.

Experiments made on top of the built dataset demonstrate potential uses through the application of machine learning approaches and also indicated the impact of the annotators' bias in the labeling process. A machine learning model aims to generalize patterns in training data to predict new data that has never been read before during the training stage. Applying ML approaches in the first set of experiments performed on top of the built dataset presents some interesting results: (1) despite the small size of the dataset, which may lead to associated challenges, such as the possibility of overfitting, that is, adapting too much to the data used in training; applying balancing methods, with the tested classifiers it was possible to reach a satisfactory F1-score (over 74% in all cases tested); (2) it was shown that even though not so robust algorithms were applied, it is possible to perform the classification solidly; (3) yet, it was shown that groups with different characteristics tend to evaluate the data differently.

Humans tend to place themselves inside one category in social groups (BODENHAUSEN; KANG; PEERY, 2012), which leads to biases towards members of their group (in-group) in terms of preferences, perception, empathy, and resource distribution (TAJFEL et al., 1971). More than exclusion, conflict, animosity, and unequal access to opportunities, the consequences of in/out-group biases profoundly contribute to the wicked challenges we currently face as a society (e.g., social injustice, extremism movements) through systemic forms of oppression, such as racism and xenophobia (TAJFEL et al., 1971). As expected, different beliefs in data labels were found throughout the descriptive data annotation procedure. As previously discussed in Section 2.2, individuals tend to organize themselves into groups based on affinity. Nevertheless, when we bring this to

politics, these affinities might be broader (e.g., as ideologies are). As noted by Müller (2019b), the left and right labels have become Brazil's preferred terminology to define those who participate more actively in political conversations. They came to bear a self-explanatory meaning as adjectives. Being on the left or on the right reveals a lot about the person who identifies itself as such. According to Müller (2019b), ideology can be understood as thought structures based on belief systems which, in turn, are defined as the gathering of those elements in which the individual deeply believes and which provide the parameters for reasoning or even for the triggering some individual and collective emotions.

In this work, we dealt with the domain of the subjective phenomena observed in the text of the tweets. When evaluating the contents present in our dataset, some might interpret some writings as endorsing or justifying intolerance. The recipient of such texts could feel hurt, outraged, or excluded, such as the one reading the text with the purpose of rate. This type of text, which can be referred to as having offensive content, represents a phenomenon not perceived equally by everyone (KOCOŃ et al., 2021). When observing the annotated data, we noticed that, in some cases where the tweet deliberately made explicit negative prejudices about people or groups, annotators signaled little or no degree of intolerance, which reinforces the influence of beliefs and ideologies over the guidelines. As put by Cover (2022), to place the everydayness of mass online hatred into an identity and ethical context requires admitting from the start that a user's subjectivity is not something that is brought to an online platform meant solely as a benign channel. Instead, online communication, like any other cultural, social, or communicative setting, actively forms, constitutes and shapes identity and belonging. Here, due to the characteristics presented in the data annotation, we see such subjectivity as being endorsed by the social stratum. The classifications indicated by different groups with common characteristics represent a reflection of individual and collective beliefs that derive from experiences lived by those in the role of labelers.

Looking at the results allows us to understand more clearly the influence of an annotator's background, and it also makes us highlight the challenges and limitations generated from this. The results achieved indicate that it is possible to detect the presence of intolerant discourses within the scope of political gender violence with the support of computational approaches. However, the result of applying such approaches is not enough to determine with certainty where there is, in fact, an intolerant speech. Since characterizing a speech as intolerant also involves recognizing the sender's intentionality, often

surrounded by subtleties when expressed, the contextual component represents a fundamental importance. Recognition of the proposed definition for the frame of symbolic violence is an important result both in recognizing that language can indeed be violent and as a way of looking at it from different perspectives.

The understanding of violence as an act that goes beyond the physical level represents a movement of understanding violence itself as a complex structure. As put by Duque (2015), cognitive strategies based on frames are sufficient to provide the necessary inputs for the construction of complex meanings and different worldviews. However, the creation of a frame of symbolic violence carries with it a series of challenges, such as how to define it, how to determine what are the core and non-core frame elements, and even how to evaluate it. In this work, we seek to define the frame of violence and its EF based on existing definitions in the literature and analogously to the available resources. However, evaluating the proposal still remains an open challenge. In this work, we evaluated the frame of symbolic violence by measuring the perception of its suitability for speeches considered intolerant by the evaluators. This evaluation approach can be understood as a still incipient method; however, which represents the validation of the popular understanding of what was proposed. We consider the lack of other forms of deeper evaluations of this frame as a limitation of this work.

Finally but not least, we remember that this work began with the adoption of a definition of what hate speech means (presented in Section 2.3), and in its final remarks, we leave here another definition, the result of a combination of proposals presented in other works (FORTUNA; NUNES, 2018; MÜLLER, 2019b) who sought to understand this broad topic. This definition does not result from the understanding that the previous one is no longer valid but rather from the understanding that just as social behavior changes dynamically, so does language and that naming and (re)defining concepts it carries is a way of understanding them and thus exploit them or even fight them when necessary. Language is alive and changing, so appropriating its different uses gives us the strength to understand our surroundings and recognize the perlocutionary power a speech can have. In concluding this work, our understanding of hate speech is as follows:

> *A language, or an indicative expression of a type of linguistic conduct, deliberately intentional, that attacks or diminishes through segregating or making explicit negative prejudices about people or groups based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other reducing their value and dignity before society,*

threatening and promoting their insecurity and, in the most extreme cases, calling for violence and extermination. A language that can occur with different linguistic styles, even in subtle forms or when humor is used.

6 CONCLUSIONS

We understand hate speech as an intersectional phenomenon capable of reaching a group of individuals not only by common isolated characteristics but also by the composition of multiple social vulnerabilities to which these individuals belong. On online communication, as in other communications forms, receivers of a speech can be crossed by intersectional discourses where one is seen as a target by features related to their race/ethnicity, gender, and social class addressed by whom produces the speech. Although there is no consensus in the literature regarding the definition of hate speech, computational methods to identify content that potentially qualifies as such have been widely studied using their own definitions. Nevertheless, as stated by KhosraviNik and Esposito (2018), despite sexist violence being a significant societal issue, both institutional and scholarly studies have frequently discounted or ignored the acknowledgment of misogyny as a manifestation of gender-based hate speech (a.k.a., misogynistic speech). Still, as delineates (RICHARDSON-SELF, 2018), the misogynistic speech "appears to illustrate all the hallmark traits of hate speech. It targets a historically and contemporary oppressed group, is characteristically hostile, systematically violent, and degrades, stigmatizes, vilifies, and disparages its targets (among other things)". Gender-based symbolic violence and discrimination in multimodal discourses are an established core in the socio-cultural regimentation of gender disparity in society (KHOSRAVINIK; ESPOSITO, 2018). As political violence, hate speech gains a legitimizing element, especially when incited by a leader and its followers. And here, we include a range of aggressive discourse intended mostly towards minorities and/or their representatives (social idea), because they are identified with (and as) the other to be eradicated. Added to the online environment effects, this kind of speech strongly reverberates (but is not limited to) on women linked to politics, where particularly extreme emotions are engaged (SĘKOWSKA-KOZŁOWSKA; BARANOWSKA; GLISZCZYŃSKA-GRABIAS, 2022).

In an environment such as social networks, where the interactions are mainly dependent on text, few computer studies observe them considering contextual and robust linguistic aspects (in addition to textual preprocessing steps) to address the hate speech detection task. Given the potential that computational approaches have in data classification, it is believed that a multidisciplinary approach capable of covering data dimensions more deeply tends to contribute to understand and to categorize them within an observation interval. Thus, when considering the significant relevance that linguistic studies

play in analyzing and understanding discourses in different media, we seek to propose an approach that uses linguistic-computational strategies to support the detection of hate speech in both areas. In this work, we focused our efforts on hate speech related to gender political violence without disregarding that the same discourses that incite hatred against women for their role in politics can and, in many cases, address intersectional spheres.

The present work addressed multiple issues related to detecting hate speech to improve the content moderation task by using natural language processing, while also focusing on bringing a semantic representation that may improve its understanding by linguistic approaches. Considering our objectives and the results achieved in the development of this study, we highlight the following contributions:

a. Proposition of a symbolic violence frame, considering a restructuring of the generic violence frame currently made available by FrameNet Brasil. In such architecture, violence is represented as physical coercion, and no other derived subframes exist. In the structure proposed in this Thesis, violence is seen as a superframe from which derives two subframes (physical violence and symbolic violence). Considering observations made on top of the results pointed out by the questionnaire respondents, it was possible to conclude that the definition proposed in this work for the symbolic violence frame is sufficient to identify more violent cases but fails to include cases of lower incidence of violence, thus needing to be revised for these cases.

b. Analysis of users' perception, observed through the completion of a questionnaire applied to users of social networks regarding their perception of the discourses conveyed in these means of communication.

c. Creation of a (manually) annotated dataset to support the identification of hate speech related to gender-based political violence covering degrees of symbolic violence present in the instances;

d. Evaluating text classification approaches both to identify linguistic characteristics associated with intolerant discourses and to classify discourse as potentially intolerant based on the presence or absence of these characteristics.

e. Validation of the "Fallacy intended to propagate hate" characteristic as a potential linguistic indicator (hallmark) of hate speech by measuring the perception of its suitability for speeches considered intolerant by evaluators through a questionnaire.

f. Contribution to the development of academic study related to social media with a focus on potentially harmful speeches.

Thus, given the above, the hypothesis raised in this work can be validated through the contributions achieved. In this way, it is true that the use of linguistic indicators that are characteristic of intolerant speeches (or even hate speech) linked to computational methods such as text classification can assist in the structuring of content disseminated in the virtual scope from the perspective of frame semantics, considering a frame of symbolic violence for that.

Thus, we summarize the impact of this work from scientific and social perspectives. From a scientific point of view, we highlight the details of the literature on the detection of hate speech, focusing mainly on methods with a more significant linguistic basis, which allowed us to identify in frame semantics a viable approach to analyze intolerant speeches. In this sense, we also highlight the proposition of the frame of symbolic violence. In addition, the definition of a reproducible protocol for experimenting with text classification methods and analyzing the results can also be understood as a contribution of a scientific nature. Also, we seek to enable the development of future studies on the part of computational linguistics under different optics. From a social point of view, this work sheds light on the lack of data on political gender violence, especially in Portuguese, and on the importance of studying this topic through practical approaches to natural language processing. However, the thesis outputs also have some limitations. Given the difficulty of getting people to label the data, the dataset built as the starting point of this study, although it represents a starting point for expanding studies in the area, consists of a small dataset. Thus, this set has limitations associated with this factor, which hinders the ability to generalize models trained from these data. Thus, it is also worth considering that, like any dataset built based on the perception of language and its interpretations, this one presents biases that are difficult to solve, given the subject's subjectivity. Another limitation of the study is related to its evaluation. This occurs due to the lack of frameworks and similar studies to be taken as a baseline for comparison.

Despite this, the activities carried out during the course of the doctorate (see Appendix F) allowed generating knowledge to help and direct future research and practical applications related to the study and detection of hate speech. In this sense, future work may explore the concept of rationales in intolerant discourses of political gender violence to investigate and define stricter patterns of these discourses. The rationale is a concept introduced by Zaidan, Eisner and Piatko (2007), which consists of portions of the texts on

which the annotator's labeling decision (e.g., intolerant or not) is based. The dataset presented here did not use rationales in its annotation process. However, due to the nature of the content, we consider the identification of rationales as a way of ascertaining stretches of a greater propensity for classifying a given content as intolerant. The use of rationales has been applied in different natural language processing proposals (MAJUMDER et al., 2021; VAFA et al., 2021). As said by Mathew et al. (2021), if these rationales are valid explanations for decisions, models that are trained to follow them might become more human-like in their decision-making process. Thus, we believe that frame elements and lexical units to the symbolical violence frame can be identified by investigating the rationales of such speeches. Another step to be taken in a future work is related to the formalization of the proposed frame in the FrameNet project. Currently, there are no means of proposing new frames to be added to the resources of the FrameNet Brasil platform, not even the Framenet of Berkeley.

REFERENCES

ABEND, O.; RAPPOPORT, A. The state of the art in semantic representation. In: **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 77–89. Available from Internet: <https://aclanthology.org/P17-1008>.

AGRAWAL, S.; AWEKAR, A. Deep learning for detecting cyberbullying across multiple social media platforms. In: PASI, G. et al. (Ed.). **Advances in Information Retrieval**. Cham: Springer International Publishing, 2018. p. 141–153. ISBN 978-3-319-76941-7.

AHMAD, W. et al. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. 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)**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 2440–2452. Available from Internet: <https://aclanthology.org/N19-1253>.

AKHTAR, S.; BASILE, V.; PATTI, V. A new measure of polarization in the annotation of hate speech. In: ALVIANO, M.; GRECO, G.; SCARCELLO, F. (Ed.). **AI*IA 2019 – Advances in Artificial Intelligence**. Cham: Springer International Publishing, 2019. p. 588–603. ISBN 978-3-030-35166-3.

ALAM, M. et al. Semantic role labeling for knowledge graph extraction from text. **Progress in Artificial Intelligence**, v. 10, n. 3, p. 309–320, Sep 2021. ISSN 2192-6360. Available from Internet: <https://doi.org/10.1007/s13748-021-00241-7>.

ALLOGHANI, M. et al. A systematic review on supervised and unsupervised machine learning algorithms for data science. In: ____. **Supervised and Unsupervised Learning for Data Science**. Cham: Springer International Publishing, 2020. p. 3–21. ISBN 978-3-030-22475-2. Available from Internet: <https://doi.org/10.1007/978-3-030-22475-2_1>.

ALMEIDA, G.; CUNHA, J. curso discurso de ódio, tô fora: ferramentas para uma internet cordial. Unpublished. 2020.

ALPAYDIN, E. **Introduction to Machine Learning**. 3. ed. London, England: MIT Press, 2014. (Adaptive Computation and Machine Learning series).

AMNESTY INTERNATIONAL. **Toxic Twitter: A toxic place for women**. 2021. [Online]. Available from Internet: <https://www.internetlab.org.br/wp-content/uploads/2021/03/5P_Relatorio_MonitorA-PT.pdf>.

AMORIM, E. et al. Brat2viz: a tool and pipeline for visualizing narratives from annotated texts. In: CEUR-WS. **Fourth Workshop on Narrative Extraction From Texts**. 2021. v. 2860, p. 49–56. Available from Internet: <https://ceur-ws.org/Vol-2860/paper6.pdf>.

ARANGO, A.; PÉREZ, J.; POBLETE, B. 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**. New York, NY, USA: Association for Computing Machinery, 2019. (SIGIR'19), p. 45–54. ISBN 9781450361729. Available from Internet: <https://doi.org/10.1145/3331184.3331262>.

ARSLAN, R. C.; WALTHER, M. P.; TATA, C. S. formr: A study framework allowing for automated feedback generation and complex longitudinal experience-sampling studies using r. **Behavior Research Methods**, Springer Science and Business Media LLC, v. 52, n. 1, p. 376–387, abr. 2019. Available from Internet: <https://doi.org/10.3758/s13428-019-01236-y>.

BADJATIYA, P. et al. Deep learning for hate speech detection in tweets. In: **Proceedings of the 26th International Conference on World Wide Web Companion**. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2017. (WWW '17 Companion), p. 759–760. ISBN 9781450349147. Available from Internet: <https://doi.org/10.1145/3041021.3054223>.

BAIDER, F. Pragmatics lost?: Overview, synthesis and proposition in defining online hate speech. **Pragmatics and Society**, John Benjamins, v. 11, n. 2, p. 196–218, 2020.

BAKER, C. F.; FILLMORE, C. J.; LOWE, J. B. The Berkeley FrameNet project. In: **36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1**. Montreal, Quebec, Canada: Association for Computational Linguistics, 1998. p. 86–90. Available from Internet: <https://aclanthology.org/P98-1013>.

BARRETT, L. et al. A lightweight yet robust approach to textual anomaly detection. In: **Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)**. Gyeongju, Republic of Korea: Association for Computational Linguistics, 2022. p. 62–67. Available from Internet: <https://aclanthology.org/2022.trac-1.8>.

BARROS, D. L. P. de. O discurso intolerante na internet: enunciação e interação. 2014.

BASSIGNANA, E.; BASILE, V.; PATTI, V. Hurtlex: A multilingual lexicon of words to hurt. In: CEUR-WS. **5th Italian Conference on Computational Linguistics, CLiC-it 2018**. 2018. v. 2253, p. 1–6. Available from Internet: <http://ceur-ws.org/Vol-2253/paper49.pdf>.

BASSO, R. **Descrição do português brasileiro**. 1st. ed. São Paulo, SP, Brasil: Párabola Editorial., 2019. (Linguística para o ensino superior). ISBN 9788579341748.

BENTÉJAC, C.; CSÖRGŐ, A.; MARTÍNEZ-MUÑOZ, G. A comparative analysis of gradient boosting algorithms. **Artificial Intelligence Review**, v. 54, n. 3, p. 1937–1967, Mar 2021. ISSN 1573-7462. Available from Internet: <https://doi.org/10.1007/s10462-020-09896-5>.

BERENBRINK, P. et al. Asynchronous opinion dynamics in social networks. In: **Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems**. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2022. (AAMAS '22), p. 109–117. ISBN 9781450392136.

BILGE, S. Théorisations féministes de l'intersectionnalité. **Diogène**, Presses Universitaires de France, n. 1, p. 70–88, 2009.

BISPO, T. D. **Arquitetura LSTM para classificação de discursos de ódio cross-lingual Inglês-PtBR**. Dissertation (Master) — Universidade Federal de Sergipe, São Cristóvão, Brasil, 2018.

BITTENCOURT, J.; FONSECA-SILVA, M. da C. Violência verbal no parlamento brasileiro: análise discursiva de um insulto e seus efeitos políticos e jurídicos / verbal violence at Brazilian's Parlament: discourse analysis of an insult and its politics and legal effects. **REVISTA DE ESTUDOS DA LINGUAGEM**, v. 28, n. 4, p. 1807–1836, 2020. ISSN 2237-2083. Available from Internet: <http://www.periodicos.letras.ufmg.br/index.php/relin/article/view/16809>.

BOAVENTURA, L. H.; FREITAS, E. C. de. Encenação e ubiquidade no twitter: a intolerância dos discursos sobre Marielle Franco. **Letrônica**, v. 13, n. 2, p. e35963, mar. 2020. Available from Internet: <https://revistaseletronicas.pucrs.br/index.php/letronica/article/view/35963>.

BODENHAUSEN, G. V.; KANG, S. K.; PEERY, D. Social categorization and the perception of social groups. **The Sage handbook of social cognition**, Sage Thousand Oaks, CA, p. 318–336, 2012.

BOURDIEU, P. Symbolic power. **Critique of Anthropology**, v. 4, n. 13-14, p. 77–85, 1979. Available from Internet: <https://doi.org/10.1177/0308275X7900401307>.

BOURDIEU, P. et al. O poder simbólico. Difel Lisboa, 1989.

BRUM, E. **Brasil, construtor de ruínas: Um olhar sobre o Brasil, de Lula a Bolsonaro**. Arquipélago Editorial, 2019. ISBN 9788554500320. Available from Internet: <https://books.google.com.br/books?id=fpO0DwAAQBAJ>.

BURCHARDT, A. et al. **Computational Semantics**. 2020. <http://www.coli.uni-saarland.de/projects/milca/courses/comsem/html/index.html>. [Online; accessed 20 August, 2020].

BUTLER, J. **Discurso de ódio: Uma política do performativo**. São Paulo: Editora Unesp, 2021.

CAMPOS, R. et al. Report on the third international workshop on narrative extraction from texts (text2story 2020). **SIGIR Forum**, Association for Computing Machinery, New York, NY, USA, v. 54, n. 1, feb. 2021. ISSN 0163-5840. Available from Internet: <https://doi.org/10.1145/3451964.3451975>.

CASTRO, L. d. R. Um estudo empírico sobre técnicas para detecção de discursos de ódio em postagens públicas escritas em português. 2019.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 785–794. ISBN 9781450342322. Available from Internet: <https://doi.org/10.1145/2939672.2939785>.

CHETTY, N.; ALATHUR, S. Hate speech review in the context of online social networks. **Aggression and Violent Behavior**, v. 40, p. 108 – 118, 2018. ISSN 1359-1789. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1359178917301064>.

CHOMSKY, N. **On Language: Chomsky's Classic Works: Language and Responsibility and Reflections on Language**. New Press, 2017. ISBN 9781595587619. Available from Internet: <https://books.google.com.br/books?id=7FzOzLgG9AkC>.

CHULVI, B.; TOSELLI, A.; ROSSO, P. **Fake News and Hate Speech: Language in Common**. arXiv, 2022. Available from Internet: <https://arxiv.org/abs/2212.02352>.

COKLUK, O. Logistic regression: Concept and application. **Educational Sciences: Theory and Practice**, ERIC, v. 10, n. 3, p. 1397–1407, 2010.

COLLINS, P.; BILGE, S. **Intersectionality**. Wiley, 2020. (Key Concepts). ISBN 9781509539680. Available from Internet: <https://books.google.com.br/books?id=5qe7yAEACAAJ>.

COLLINS, P. H. On violence, intersectionality and transversal politics. **Ethnic and Racial Studies**, Taylor & Francis, v. 40, n. 9, p. 1460–1473, 2017.

COVER, R. Digital hostility, subjectivity and ethics: Theorising the disruption of identity in instances of mass online abuse and hate speech. **Convergence**, v. 0, n. 0, p. 1–14, 2022. Available from Internet: <https://doi.org/10.1177/13548565221122908>.

DADICO, C. M. **O Ódio Ancestral Como Elemento Constitutivo Do Estado Moderno e Seus Reflexos Na Compreensão dos Crimes De Ódio: Um Diálogo Entre o Direito Internacional e o Direito Brasileiro**. Thesis (PhD) — Programa de Pós-Grduação em Ciências Criminais da Escola de Direito da Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, RS, Brazil, 2020.

DAVIDSON, T. et al. Automated hate speech detection and the problem of offensive language. **Proceedings of the International AAAI Conference on Web and Social Media**, v. 11, n. 1, May 2017. Available from Internet: <https://ojs.aaai.org/index.php/ICWSM/article/view/14955>.

DOMINGOS, P. A few useful things to know about machine learning. **Communications of the ACM**, Association for Computing Machinery (ACM), v. 55, n. 10, p. 78–87, oct. 2012. Available from Internet: <https://doi.org/10.1145/2347736.2347755>.

DUNKER, C. et al. **Ética e pós-verdade**. Porto Alegre: Editora Dublinense, 2018.

DUQUE, P. H. Discurso e cognição: uma abordagem baseada em frames. **Revista da ANPOLL**, v. 1, n. 39, p. 25–48, 2015.

ELSHERIEF, M. et al. **Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media**. 2018.

ERK, K. **Computational Semantics**. Oxford University Press, 2018. Available from Internet: <https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-331>.

ESRA'M, A. Lexicon-based detection of violence on social media. **Cognitive Semantics**, Brill, v. 5, n. 1, p. 32–69, 2019.

EVANS, V.; GREEN, M. **Cognitive Linguistics: An Introduction**. L. Erlbaum, 2006. ISBN 9780805860146. Available from Internet: <https://books.google.com.br/books?id=vrafVlXvFmcC>.

FACELI, K. et al. **Inteligência artificial: uma abordagem de aprendizado de máquina**. Grupo Gen - LTC, 2011. ISBN 9788521618805. Available from Internet: <https://books.google.com.br/books?id=4DwelAEACAAJ>.

FILLMORE, C. J. Linguistics in the morning calm. **Seoul: Hanshin Publishing Co**, p. 111–137, 1982.

FILLMORE, C. J. Frames and the semantics of understanding. **Quaderni di semantica**, v. 6, n. 2, p. 222–254, 1985.

FILLMORE, C. J. Valency issues in framenet. **TRENDS IN LINGUISTICS STUDIES AND MONOGRAPHS**, Mouton de Gruyter, v. 187, p. 129, 2007.

FILLMORE, C. J. et al. Frame semantics. **Cognitive linguistics: Basic readings**, Mouton de Gruyter New York, v. 34, p. 373–400, 2006.

FORTUNA, P. **Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes**. Dissertation (Master) — Universidade do Porto, Porto, Portugal, 2017.

FORTUNA, P.; NUNES, S. A survey on automatic detection of hate speech in text. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 51, n. 4, p. 1–30, 2018.

FRAGOSO, S.; AMARAL, A.; RECUERO, R. **MÉTODOS DE PESQUISA PARA INTERNET**. SULINA, 2011. ISBN 9788520505946. Available from Internet: <https://books.google.com.br/books?id=utWluAAACAAJ>.

FREITAS, C. **Linguística Computacional**. 1st. ed. São Paulo, SP, Brasil: Parábola, 2022. (Linguística para o ensino superior).

FU, S. et al. Social media overload, exhaustion, and use discontinuance: Examining the effects of information overload, system feature overload, and social overload. **Information Processing & Management**, v. 57, n. 6, p. 102307, 2020. ISSN 0306-4573. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0306457320308025>.

GAO, L.; HUANG, R. **Detecting Online Hate Speech Using Context Aware Models**. 2018.

GILLESPIE, T. **Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media**. Yale University Press, 2018. ISBN 9780300235029. Available from Internet: <https://books.google.com.br/books?id=cOJgDwAAQBAJ>.

GLUCKSMANN, A. El discurso del odio. **Desde el Jardín de Freud**, v. 19, p. 328–333, 2019.

GRÖNDAHL, T. et al. All You Need is "Love": Evading Hate Speech Detection. In: **Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security**. New York, NY, USA: Association for Computing Machinery, 2018. (AISec '18), p. 2–12. ISBN 9781450360043. Available from Internet: <https://doi.org/10.1145/3270101.3270103>.

GRUS, J. **Data Science from Scratch**. Sebastopol, CA: O'Reilly Media, 2015.

HANKIVSKY, O. Intersectionality 101. **The Institute for Intersectionality Research & Policy, SFU**, p. 1–34, 2014.

HE, G. et al. Think beyond the word: Understanding the implied textual meaning by digesting context, local, and noise. In: **Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: Association for Computing Machinery, 2020. (SIGIR '20), p. 2297–2306. ISBN 9781450380164. Available from Internet: <https://doi.org/10.1145/3397271.3401435>.

HE, H.; GARCIA, E. A. Learning from imbalanced data. **IEEE Transactions on knowledge and data engineering**, Ieee, v. 21, n. 9, p. 1263–1284, 2009.

HOFMANN, T. Unsupervised learning by probabilistic latent semantic analysis. **Machine learning**, Springer Nature BV, v. 42, n. 1-2, p. 177, 2001.

HUANG, Y. **The Oxford Handbook of Pragmatics**. Oxford University Press, 2017. (Oxford handbooks in linguistics). ISBN 9780199697960. Available from Internet: <https://books.google.pt/books?id=PlvjDQAAQBAJ>.

IBGE. **Desigualdades sociais por cor ou raça no Brasil**. Av. Franklin Roosevelt, 166 - Centro - 20021-120 - Rio de Janeiro, RJ - Brasil: IBGE, 2019. Estudos e Pesquisas Informação Demográfica e Socioeconômica número 41. ISBN 9788524045134.

ISA, D.; HIMELBOIM, I. A social networks approach to online social movement: Social mediators and mediated content in #freeajstaff twitter network. **Social Media + Society**, v. 4, n. 1, p. 2056305118760807, 2018. Available from Internet: <https://doi.org/10.1177/2056305118760807>.

ISRAELI, A.; TSUR, O. Free speech or free hate speech? analyzing the proliferation of hate speech in parler. In: **Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)**. Seattle, Washington (Hybrid): Association for Computational Linguistics, 2022. p. 109–121. Available from Internet: <https://aclanthology.org/2022.woah-1.11>.

JAHAN, M. S.; OUSSALAH, M. **A systematic review of Hate Speech automatic detection using Natural Language Processing**. arXiv, 2021. Available from Internet: <https://arxiv.org/abs/2106.00742>.

JAPKOWICZ, N.; SHAH, M. **Evaluating Learning Algorithms: A Classification Perspective**. Cambridge: Cambridge University Press, 2011.

JIANG, S.; ROBERTSON, R. E.; WILSON, C. Reasoning about political bias in content moderation. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 34, n. 09, p. 13669–13672, Apr 2020. Available from Internet: <https://ojs.aaai.org/index.php/AAAI/article/view/7117>.

JIAWEN, D. et al. Cold: A benchmark for chinese offensive language detection. **ArXiv**, abs/2201.06025, 2022.

JURAFSKY, D. Pragmatics and computational linguistics. In: **The Handbook of Pragmatics**. Blackwell Publishing Ltd, 2004. p. 578–604. Available from Internet: <https://doi.org/10.1002/9780470756959.ch26>.

KAYE, D. A. **Speech police: The global struggle to govern the Internet**. 91 Claremont Avenue, Suite 515. New York, NY 10027: Columbia Global Reports, 2019.

KE, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. **Advances in neural information processing systems**, v. 30, 2017.

KHOSRAVINIK, M.; ESPOSITO, E. Online hate, digital discourse and critique: Exploring digitally-mediated discursive practices of gender-based hostility. **Lodz Papers in Pragmatics**, v. 14, n. 1, p. 45–68, 2018. Available from Internet: <https://doi.org/10.1515/lpp-2018-0003>.

KHURANA, U. et al. Hate speech criteria: A modular approach to task-specific hate speech definitions. In: **Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)**. Seattle, Washington (Hybrid): Association for Computational Linguistics, 2022. p. 176–191. Available from Internet: <https://aclanthology.org/2022.woah-1.17>.

KOCOŃ, J. et al. Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. **Information Processing & Management**, v. 58, n. 5, p. 102643, 2021. ISSN 0306-4573. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0306457321001333>.

KOTSIANTIS, S. B. et al. Supervised machine learning: A review of classification techniques. **Emerging artificial intelligence applications in computer engineering**, Amsterdam, v. 160, n. 1, p. 3–24, 2007.

KRIPPENDORFF, K. **Content analysis**. 2. ed. Thousand Oaks, CA: SAGE Publications, 2004.

LAKOFF, G.; DEAN, H.; HAZEN, D. **Don't Think of an Elephant!: Know Your Values and Frame the Debate : the Essential Guide for Progressives**. Chelsea Green Publishing Company, 2004. (New York Times Bestseller). ISBN 9781931498715. Available from Internet: <https://books.google.com.br/books?id=dovUAgAAQBAJ>.

LAVALLEY, M. P. Logistic regression. **Circulation**, Am Heart Assoc, v. 117, n. 18, p. 2395–2399, 2008.

LEES, A. et al. **A New Generation of Perspective API: Efficient Multilingual Character-level Transformers**. arXiv, 2022. Available from Internet: <https://arxiv.org/abs/2202.11176>.

LI, C.; SUN, A.; DATTA, A. Twevent: Segment-based event detection from tweets. In: **Proceedings of the 21st ACM International Conference on Information and Knowledge Management**. New York, NY, USA: Association for Computing Machinery, 2012. (CIKM '12), p. 155–164. ISBN 9781450311564. Available from Internet: <https://doi.org/10.1145/2396761.2396785>.

LI, M. et al. Twitter as a tool for social movement: An analysis of feminist activism on social media communities. **Journal of community psychology**, Wiley Online Library, v. 49, n. 3, p. 854–868, 2021.

LIBBRECHT, M. W.; NOBLE, W. S. Machine learning applications in genetics and genomics. **Nature Reviews Genetics**, Nature Publishing Group UK London, v. 16, n. 6, p. 321–332, 2015.

LIU, Z.; LIN, Y.; SUN, M. **Representation Learning for Natural Language Processing**. Springer Singapore, 2020. Available from Internet: <https://doi.org/10.1007\%2F978-981-15-5573-2>.

MA, Y.; HE, H. Imbalanced learning: foundations, algorithms, and applications. John Wiley & Sons, 2013.

MACAVANEY, S. et al. Hate speech detection: Challenges and solutions. **PloS one**, Public Library of Science, v. 14, n. 8, 2019.

MAJUMDER, B. P. et al. **Rationale-Inspired Natural Language Explanations with Commonsense**. arXiv, 2021. Available from Internet: <https://arxiv.org/abs/2106.13876>.

MANI, I.; ZHANG, I. knn approach to unbalanced data distributions: a case study involving information extraction. In: ICML. **Proceedings of workshop on learning from imbalanced datasets**. [S.l.], 2003. v. 126, p. 1–7.

MATHEW, B. et al. Hatexplain: A benchmark dataset for explainable hate speech detection. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 35, n. 17, p. 14867–14875, May 2021. Available from Internet: <https://ojs.aaai.org/index.php/AAAI/article/view/17745>.

MATTHES, J. et al. "too much to handle": Impact of mobile social networking sites on information overload, depressive symptoms, and well-being. **Computers in Human Behavior**, v. 105, p. 106217, 2020. ISSN 0747-5632. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0747563219304364>.

MAURO, T. D. Le parole per ferire. **Internazionale**, v. 27, n. 9, p. 2016, 2016.

MAYNARD, J. L.; BENESCH, S. Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention. **Genocide Studies and Prevention**, University of Toronto Press, 2016.

MINSKY, M. **A framework for representing knowledge**. MIT, Cambridge, 1974. Available from Internet: <https://courses.media.mit.edu/2004spring/mas966/Minsky\%201974%20Framework%20for%20knowledge.pdf>.

MITCHELL, T. **Machine Learning**. New York, NY: McGraw-Hill Professional, 1997. (McGraw-Hill series in computer science).

MITKOV, R. **The Oxford Handbook of Computational Linguistics**. OUP Oxford, 2004. (Oxford Handbooks Series). ISBN 9780199276349. Available from Internet: <https://books.google.com.br/books?id=yl6AnaKtVAkC>.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. **Sistemas inteligentes-Fundamentos e aplicações**, v. 1, n. 1, p. 32, 2003.

MÜLLER, A. **Política do ódio no Brasil**. Viseu, 2019. ISBN 9788530012953. Available from Internet: <https://books.google.com.br/books?id=9PO-DwAAQBAJ>.

MÜLLER, A. A. C. **Brasil polarizado : os discursos de incitação ao ódio na campanha presidencial de 2014**. Thesis (PhD) — Programa de Pós-Graduação em Comunicação Social, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, RS, Brazil, 2019. Available from Internet: <https://tede2.pucrs.br/tede2/handle/tede/8720>.

MURPHY, K. P. et al. Naive bayes classifiers. **University of British Columbia**, v. 18, n. 60, p. 1–8, 2006.

NAGARAJAN, M. et al. Spatio-temporal-thematic analysis of citizen sensor data: Challenges and experiences. In: VOSSEN, G.; LONG, D. D. E.; YU, J. X. (Ed.). **Web Information Systems Engineering - WISE 2009**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. p. 539–553. ISBN 978-3-642-04409-0.

NEMER, D. The three types of Whatsapp users getting Brazil's Jair Bolsonaro elected. **The Guardian**, v. 25, 2018.

OFOGHI, B.; YEARWOOD, J.; GHOSH, R. A semantic approach to boost passage retrieval effectiveness for question answering. In: **Proceedings of the 29th Australasian Computer Science Conference - Volume 48**. AUS: Australian Computer Society, Inc., 2006. (ACSC '06), p. 95–101. ISBN 1920682309.

OSSWALD, R.; VALIN, R. D. V. Framenet, frame structure, and the syntax-semantics interface. In: ____. **Frames and Concept Types: Applications in Language and Philosophy**. Cham: Springer International Publishing, 2014. p. 125–156. ISBN 978-3-319-01541-5. Available from Internet: <https://doi.org/10.1007/978-3-319-01541-5_6>.

PAIVA, P.; SILVA, V. da; MOURA, R. Detecção automática de discurso de ódio em comentários online. In: **Anais da VII Escola Regional de Computação Aplicada à Saúde**. Porto Alegre, RS, Brasil: SBC, 2019. p. 157–162. Available from Internet: <https://sol.sbc.org.br/index.php/ercas/article/view/9052>.

PAIVA, P. D.; SILVA, V. Matias da; MOURA, R. S. Detecção de discurso de ódio utilizando vetores de features aplicados a uma base nova de comentários em português. **Revista de Sistemas e Computação-RSC**, v. 10, n. 1, 2020.

PALEYES, A.; URMA, R.-G.; LAWRENCE, N. D. Challenges in deploying machine learning: a survey of case studies. **ACM Computing Surveys**, ACM New York, NY, v. 55, n. 6, p. 1–29, 2022.

PASQUALI, A. et al. TLS-covid19: A new annotated corpus for timeline summarization. In: **Lecture Notes in Computer Science**. Springer International Publishing, 2021. p. 497–512. Available from Internet: <https://doi.org/10.1007/978-3-030-72113-8_33>.

PATEL, A.; MEEHAN, K. Fake News Detection on Reddit utilising Countvectorizer and Term Frequency-Inverse Document Frequency with Logistic Regression, MultinominalNB and Support Vector Machine. In: **2021 32nd Irish Signals and Systems Conference (ISSC)**. [S.l.: s.n.], 2021. p. 1–6.

PAVLOPOULOS, J.; MALAKASIOTIS, P.; ANDROUTSOPOULOS, I. Deep learning for user comment moderation. In: **Proceedings of the First Workshop on Abusive Language Online**. Vancouver, BC, Canada: Association for Computational Linguistics, 2017. p. 25–35. Available from Internet: <https://www.aclweb.org/anthology/W17-3004>.

PELLE, R. de; MOREIRA, V. Offensive comments in the brazilian web: a dataset and baseline results. In: **Anais do VI Brazilian Workshop on Social Network Analysis and Mining**. Porto Alegre, RS, Brasil: SBC, 2017. ISSN 2595-6094. Available from Internet: <https://sol.sbc.org.br/index.php/brasnam/article/view/3260>.

PÉREZ, J. M.; GIUDICI, J. C.; LUQUE, F. M. pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks. **CoRR**, abs/2106.09462, 2021. Available from Internet: <https://arxiv.org/abs/2106.09462>.

PETRUCK, M. R. L.; ELLSWORTH, M. J. Representing spatial relations in FrameNet. In: **Proceedings of the First International Workshop on Spatial Language Understanding**. New Orleans: Association for Computational Linguistics, 2018. p. 41–45. Available from Internet: <https://www.aclweb.org/anthology/W18-1405>.

PIAO, B. et al. Real-time event detection and tracking in microblog via text chain and sentiment time series. In: **2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC)**. [S.l.: s.n.], 2020. p. 178–185.

PINHO, T. R. d. Debaixo do Tapete: A Violência Política de Gênero e o Silêncio do Conselho de Ética da Câmara dos Deputados. **Revista Estudos Feministas**, scielo, v. 28, 00 2020. ISSN 0104-026X. Available from Internet: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-026X2020000200202&nrm=iso>.

POLGUÈRE, A.; MEL'ČUK, I. **Dependency in Linguistic Description**. John Benjamins Publishing Company, 2009. (Studies in language companion series). ISBN 9789027205780. Available from Internet: <https://books.google.nl/books?id=72I2AegupkYC>.

POPPER, K. **The open society and its enemies**. Princeton, NJ: Princeton University Press, 2013.

RECUERO, R.; SOARES, F. B.; GRUZD, A. Hyperpartisanship, Disinformation and Political Conversations on Twitter: The Brazilian Presidential Election of 2018. **Proceedings of the International AAAI Conference on Web and Social Media**, v. 14, n. 1, p. 569–578, May 2020. Available from Internet: <https://ojs.aaai.org/index.php/ICWSM/article/view/7324>.

REVISTA AZMINA; INTERNETLAB. **MonitorA: relatório sobre violência política online em páginas e perfis de candidatas(os) nas eleições municipais de 2020**. 2021. [Online] São Paulo. Available from Internet: <https://www.internetlab.org.br/wp-content/uploads/2021/03/5P_Relatorio_MonitorA-PT.pdf>.

RICHARDSON-SELF, L. Woman-hating: On misogyny, sexism, and hate speech. **Hypatia**, Wiley Online Library, v. 33, n. 2, p. 256–272, 2018.

ROBERTS, S. T. **Behind the Screen**. Yale University Press, 2017. Available from Internet: <https://doi.org/10.2307/j.ctvhrcz0v>.

RODRIGUEZ, J. D.; PEREZ, A.; LOZANO, J. A. Sensitivity analysis of k-fold cross validation in prediction error estimation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 32, n. 3, p. 569–575, March 2010. ISSN 0162-8828.

ROSA, J.; BURDICK, C. Language ideologies. In: GARCÍA, O.; FLORES, N.; SPOTTI, M. (Ed.). **Language and Society**. Oxford: Oxford University Press, 2016. p. 103–123. Available from Internet: <https://doi.org/10.1093/oxfordhb/9780190212896.013.15>.

RÖTTGER, P. et al. Two contrasting data annotation paradigms for subjective NLP tasks. **CoRR**, abs/2112.07475, 2021. Available from Internet: <https://arxiv.org/abs/2112.07475>.

RÖTTGER, P. et al. HateCheck: Functional tests for hate speech detection models. 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)**. Online: Association for Computational Linguistics, 2021. p. 41–58. Available from Internet: <https://aclanthology.org/2021.acl-long.4>.

RUAS, T. et al. Enhanced word embeddings using multi-semantic representation through lexical chains. **Information Sciences**, v. 532, p. 16 – 32, 2020. ISSN 0020-0255. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0020025520303911>.

RUPPENHOFER, J. et al. **FrameNet II: Extended theory and practice**. Berkeley, CA, USA, 2016. Available from Internet: <https://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf>.

SALOMÃO, M. M. M.; TORRENT, T. T.; SAMPAIO, T. F. A linguística cognitiva encontra a linguística computacional: notícias do projeto framenet brasil. **Cadernos de Estudos Lingüísticos**, v. 55, n. 1, p. 7–34, 2013.

SANTANA, B. et al. A survey on narrative extraction from textual data. **Artificial Intelligence Review**, Jan 2023. ISSN 1573-7462. Available from Internet: <https://doi.org/10.1007/s10462-022-10338-7>.

SANTANA, B. S.; VANIN, A. A. Detecting group beliefs related to 2018's brazilian elections in tweets: A combined study on modeling topics and sentiment analysis. In: **Proceedings of the Workshop on Digital Humanities and Natural Language Processing (DHandNLP 2020) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2020)**. [s.n.], 2020. Available from Internet: <https://ceur-ws.org/Vol-2607/paper2.pdf>.

SANTANA, B. S.; VANIN, A. A.; WIVES, L. K. Sexist hate speech: Identifying potential online verbal violence instances. In: PINHEIRO, V. et al. (Ed.). **Computational Processing of the Portuguese Language**. Cham: Springer International Publishing, 2022. p. 177–187. ISBN 978-3-030-98305-5.

SARDENBERG, C. M. A violência simbólica de gênero e a lei "antibaixaria" na Bahia. **OBSERVE: NEIM/UFBA**, 2011.

SCHIRMER, L. C.; DALMOLIN, A. R. Discurso de ódio biopolítico no caso Marielle Franco. In: **I Congresso Nacional de Biopolítica e Direitos Humanos**. [s.n.], 2018. Available from Internet: <https://publicacoeseventos.unijui.edu.br/index.php/conabipodihu/article/view/9294/7960>.

SCHOMACKER, T.; TROPMANN-FRICK, M. Language representation models: An overview. **Entropy**, v. 23, n. 11, 2021. ISSN 1099-4300. Available from Internet: <https://www.mdpi.com/1099-4300/23/11/1422>.

SCHUBERT, L. Semantic representation. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 29, n. 1, Mar 2015. Available from Internet: <https://doi.org/10.1609/aaai.v29i1.9759>.

SCHUBERT, L. Computational linguistics. In: ZALTA, E. N. (Ed.). **The Stanford Encyclopedia of Philosophy**. Spring 2020. [S.l.]: Metaphysics Research Lab, Stanford University, 2020.

SCOTT, J. W. Gender: a useful category of historical analysis. **The American historical review**, JSTOR, v. 91, n. 5, p. 1053–1075, 1986.

SĘKOWSKA-KOZŁOWSKA, K.; BARANOWSKA, G.; GLISZCZYŃSKA-GRABIAS, A. Sexist hate speech and the international human rights law: Towards legal recognition of the phenomenon by the united nations and the council of europe. **International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique**, v. 35, n. 6, p. 2323–2345, Dec 2022. ISSN 1572-8722. Available from Internet: <https://doi.org/10.1007/s11196-022-09884-8>.

SILVA, L. A. et al. Analyzing the targets of hate in online social media. **CoRR**, abs/1603.07709, 2016. Available from Internet: <http://arxiv.org/abs/1603.07709>.

SILVA, P. H. da. **De Louca a Incompetente: Construções Discursivas Em Relação à Ex-Presidenta Dilma Rousseff**. Thesis (PhD) — Programa de Pós-Grduação em Estudos de Linguagem da Universidade Federal de Mato Grosso, 2019.

SILVA, S.; SERAPIAO, A. Detecção de discurso de ódio em português usando cnn combinada a vetores de palavras. In: **Proceedings of KDMILE 2018, Symposium on Knowledge Discovery, Mining and Learning, São Paulo, SP, Brazil**. [S.l.: s.n.], 2018.

SINGH, G. et al. Comparison between multinomial and bernoulli naïve bayes for text classification. In: IEEE. **2019 International Conference on Automation, Computational and Technology Management (ICACTM)**. [S.l.], 2019. p. 593–596.

SONG, Y.-Y.; YING, L. Decision tree methods: applications for classification and prediction. **Shanghai archives of psychiatry**, Shanghai Mental Health Center, v. 27, n. 2, p. 130, 2015.

SOUSA, J. G. R. de. **Feature extraction and selection for automatic hate speech detection on Twitter**. Dissertation (Master) — Faculdade de Engenharia da Universidade do Porto, Porto, Portugal, 2019.

TAJFEL, H. et al. Social categorization and intergroup behaviour. **European journal of social psychology**, Wiley Online Library, v. 1, n. 2, p. 149–178, 1971.

TAY, Y. et al. Reasoning with sarcasm by reading in-between. In: **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Melbourne, Australia: Association for Computational Linguistics, 2018. p. 1010–1020. Available from Internet: <https://www.aclweb.org/anthology/P18-1093>.

TEIXEIRA, S. H.; ZAMORA, M. H. Pensando a interseccionalidade a partir da vida e morte de Marielle Franco. **Dignidade Re-Vista**, v. 4, n. 7, p. 139–153, 2019.

Terra de Direitos e a Justiça Global. **Violência Política e Eleitoral no Brasil: Panorama das violações de direitos humanos de 2016 a 2020**. 2020. Avaible at: <http://www.global.org.br/wp-content/uploads/2020/09/Relatório_Violencia-Politica_FN.pdf>.

THAPAR-BJÖRKERT, S.; SAMELIUS, L.; SANGHERA, G. S. Exploring symbolic violence in the everyday: misrecognition, condescension, consent and complicity. **Feminist review**, SAGE Publications Sage UK: London, England, v. 112, n. 1, p. 144–162, 2016.

TIRRELL, L. Toxic speech: Toward an epidemiology of discursive harm. **philosophical topics**, JSTOR, v. 45, n. 2, p. 139–162, 2017.

TORRENT, T. T.; ELLSWORTH, M. Behind the labels: Criteria for defining analytical categories in framenet brasil. **Veredas-Revista de Estudos Linguisticos**, Federal University of Juiz de Fora (UFJF), v. 17, n. 1, p. 44–66, 2013.

TORRENT, T. T. et al. Representing context in FrameNet: A multidimensional, multimodal approach. **Frontiers in Psychology**, Frontiers Media SA, v. 13, abr. 2022. Available from Internet: <https://doi.org/10.3389/fpsyg.2022.838441>.

VAFA, K. et al. Rationales for sequential predictions. In: **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. p. 10314–10332. Available from Internet: <https://aclanthology.org/2021.emnlp-main.807>.

WALDRON, J. **The harm in hate speech**. 79 Garden Street, Cambridge, MA 02138, USA: Harvard University Press, 2012.

WAZLAWICK, R. S. **Metodologia de Pesquisa para Ciência da Computação**. 3rd. ed. Barueri, SP, Brasil: GEN LTC, 2020.

WICKRAMASINGHE, I.; KALUTARAGE, H. Naive bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. **Soft Computing**, Springer, v. 25, n. 3, p. 2277–2293, 2021.

WILSON, R. A.; LAND, M. K. Hate speech on social media: Towards a context-specific content moderation policy. **Connecticut Law Review, Forthcoming**, 2020.

YANG, K.; JANG, W.; CHO, W. I. **APEACH: Attacking Pejorative Expressions with Analysis on Crowd-Generated Hate Speech Evaluation Datasets**. arXiv, 2022. Available from Internet: <https://arxiv.org/abs/2202.12459>.

YEN, S.-J.; LEE, Y.-S. Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. In: HUANG, D.-S.; LI, K.; IRWIN, G. W. (Ed.). **Intelligent Control and Automation: International Conference on Intelligent Computing, ICIC 2006 Kunming, China, August 16–19, 2006**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. p. 731–740. ISBN 978-3-540-37256-1. Available from Internet: <https://doi.org/10.1007/978-3-540-37256-1_89>.

ZAIDAN, O.; EISNER, J.; PIATKO, C. Using “annotator rationales” to improve machine learning for text categorization. In: **Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference**. Rochester, New York: Association for Computational Linguistics, 2007. p. 260–267. Available from Internet: <https://aclanthology.org/N07-1033>.

ZHANG, Y.; JIN, R.; ZHOU, Z.-H. Understanding bag-of-words model: a statistical framework. **International Journal of Machine Learning and Cybernetics**, v. 1, n. 1, p. 43–52, Dec 2010. ISSN 1868-808X. Available from Internet: <https://doi.org/10.1007/s13042-010-0001-0>.

ZHENG, C.; WANG, Y.; CHANG, B. **Query Your Model with Definitions in FrameNet: An Effective Method for Frame Semantic Role Labeling**. arXiv, 2022. Available from Internet: <https://arxiv.org/abs/2212.02036>.

ZOBEL, J. **Writing for Computer Science**. Springer London, 2014. Available from Internet: <https://doi.org/10.1007/978-1-4471-6639-9>.

APPENDIX A — RESUMO EXPANDIDO

Uma Abordagem Linguístico-Computacional para Auxiliar a Análise da Configuração Discursiva da Violência nas Redes Sociais

Nas últimas décadas, a tecnologia da informação evoluiu enormemente, com uma expressiva adoção de redes sociais online e plataformas de mídia social. A tendência simbiótica contínua em relação ao aumento do consumo de informações eletrônicas e produção de dados por usuários finais usando esses sistemas eletrônicos indica a análise de tais dados como uma área próspera de pesquisa e desenvolvimento contínuos, com novos recursos sendo criados diariamente (dados estruturados e não estruturados). Esse aumento revolucionou a comunicação ao permitir uma interação digital rápida, fácil e quase gratuita entre seus usuários. Entretanto, em meio a tanto conteúdo disseminado, discursos violentos direcionados a grupos minorizados ganharam força, muitas vezes se valendo de questões como "anonimidade", falta de moderação e uma interpretação equivocada dos limites da liberdade de expressão. As pesquisas destinadas a estudar o discurso de ódio, principalmente em sua forma online, cresceram nos últimos anos; no entanto, as abordagens capazes de detectar automaticamente esse tipo de conteúdo ainda apresentam limitações significativas. Essas limitações são ainda mais latentes em línguas com dados escassos, como a língua portuguesa.

Este trabalho se propõe ao estudo do uso de indicadores linguísticos característicos a discursos de ódio associados a métodos computacionais (como a classificação de textos por algoritmos de aprendizado de máquina). Assim, buscamos avaliar o uso de tais indicadores como forma de viabilização de enquadramento de conteúdos disseminados em português brasileiro em redes sociais sob a ótica da Semântica de Frames, considerando a instanciação de um frame de violência simbólica como uma forma de propor um meio de interposição de discursos intolerantes. Discursos de ódio são disseminados em diversas esferas e têm como alvo diferentes grupos, sendo a forma de atacar aqueles vistos como alvos diferentes a cada contexto. No escopo deste trabalho, buscamos identificar o discurso de ódio relacionado à violência política de gênero cobrindo diferentes graus de violência simbólica presentes nas instâncias analisadas.

Como fonte primária de dados, consideramos a extração de dados do Twitter, de modo que a coleta abarque termos considerados relevantes para pesquisa, isto é, que apresentem tópicos tidos como potencialmente portadores de discurso de ódio. Ao tratar dos dados a serem observados, propõe-se ranqueamento de textos coletados através do

uso da API Perspective. De forma a analisá-los, criamos um dataset de tweets de contexto político manualmente anotados. A partir deste conjunto de dados, pudemos validar o uso do frame proposto de violência simbólica como uma forma de representar discursos avaliados com um maior grau de intolerância. Também a partir deste conjunto de dados realizamos uma série de experimentos de classificação com o intuito de identificar a presença de características intolerantes associadas a discursos de ódio, e a partir dessa identificação classificar potenciais tweets com discurso intolerante. Para realização da análise dos dados envolvidos na pesquisa, consideramos uma abordagem quali-quantitativa na qual, pelo aspecto qualitativo, em um primeiro momento consideramos uma análise de conteúdo utilizando abordagens computacionais (por exemplo, uso da API Perspective para seleção de dados, algoritmos de classificação de textos), a qual servirá de base para uma análise qualitativa das observações, e ainda das questões abertas em questionário. Já a abordagem quantitativa faz uso de análises estatísticas, como distribuições de frequência, correlações e representações gráficas, medidas de dispersão, medidas de tendência central, buscando assim observar como, e se, a presente proposta se adéqua aos desfechos esperados. A partir do uso desta abordagem, os desfechos deste trabalho nos levam apontar contribuições tanto de impacto científico quanto social, com os quais buscamos enriquecer o desenvolvimento estudos centrados em redes sociais, com foco em discursos potencialmente intolerantes escritos em português, levando também a percepção do considerando sobre o conteúdo gerado nas redes sociais e suas repercussões no cotidiano.

Dentre as principais contribuições deste trabalho destacamos:

a. Proposição de um frame de violência simbólica, considerando uma reestruturação do enquadramento genérico da violência disponibilizado atualmente pelo FrameNet Brasil. Nessa arquitetura, a violência é representada como coerção física e não existem outros subframes derivados. Na estrutura proposta, a violência é vista como um superframe, do qual derivam dois subframes (violência física e violência simbólica). Considerando as observações feitas sobre os resultados apontados pelos respondentes do questionário, foi possível concluir que a definição proposta neste trabalho para o quadro de violência simbólica é suficiente para identificar casos mais violentos e não inclui casos de menor incidência de violência, portanto, precisam ser revistos para esses casos.

b. Análise da percepção dos usuários, observada por meio do preenchimento de um questionário aplicado a usuários de redes sociais sobre sua percepção sobre os discursos veiculados em português brasileiro nesses meios de comunicação.

c. Construção de um conjunto de dados anotados manualmente para apoiar a identificação do discurso de ódio relacionado à violência política baseada em gênero, abrangendo os graus de violência simbólica presentes nas instâncias;

d. A experimentação com classificação de texto aborda tanto a identificação de características linguísticas associadas a discursos intolerantes quanto a classificação do discurso como potencialmente intolerante com base na presença ou ausência dessas características.

e. Validação da característica "Falácia com intenção de propagar ódio" como um potencial indicador linguístico do discurso de ódio através da aferição, por meio de questionário, da percepção de sua adequação a discursos considerados intolerantes por avaliadores.

f. Contribuição para o desenvolvimento de um estudo acadêmico centrado nas redes sociais com enfoque em discursos potencialmente lesivos; e a reflexão do usuário sobre os comportamentos apresentados nessas mídias e suas repercussões no cotidiano.

APPENDIX B — FRAME DE VIOLÊNCIA SIMBÓLICA

DEFINIÇÃO

- Este frame descreve atos (ou situações caracterizadas por atos) que resultam em uma forma de violência sem coerção física, isto é, violência simbólica, causando danos morais e psicológicos. Os atos podem envolver um Agressor ferindo uma Vítima, ou ainda Agressores causando danos simbólicos uns aos outros.
  - Eles cometeram várias difamações em relação à vítima do caso.
  - As manifestações preconceituosas feitas em seu discurso proferido na semana passada tiveram grande repercussão.

ELEMENTOS DE FRAME NUCLEARES

- Agressor [Aggressor] o Agressor é a pessoa que causa dano à vítima.
  - Suas ameaças para com seus concorrentes políticos parece não conhecer limites.
- Agressores [Aggressors] os agressores cometem atos de violência em conjunto
  - Ataques coordenados por instituições em mídias virtuais tornaram-se uma estratégia para desmoralizar pessoas e/ou entidades.
- Vítima [Victim] ser ou entidade que sofre o dano;
  - O aumento da violência por meios virtuais contra opositores da proposta do governo atrai a atenção internacional.

ELEMENTOS DE FRAME NÃO NUCLEARES

- Circunstâncias [Circumstances] Circunstâncias descreve uma situação (em um tempo e lugar particulares) que é especificamente independente do ato violento ou de qualquer um de seus participantes;
- Interpretação [Interpretation] Ação ou característica atribuída à vítima cujo agressor endereça em ato simbolicamente violento (e.g., por meio de atitudes e/ou discursos referindo à vítima);
- Evento_continente [Containing_event] Identifica o evento no qual o dano foi causado.

- Finalidade [Purpose] Identifica a Finalidade pela qual a ação que causa o dano é realizada.
  - semantic_type: @state_of_affairs
- Frequência [Frequency] Com que Frequência a violência ocorre.
  - Figuras de oposição são diariamente alvo de atos violentos em suas redes sociais onde busca-se descredibilizar sua imagem
- Grau [Degree] Indica o Grau da violência cometida
  - semantic_type: @degree
  - As campanhas eleitorais têm sido palco de extrema disseminação do ódio
- Iterações [Iterations] Iterações faz referência ao número de vezes que o ato violento ocorre.
- Meios [Means] Meios utilizados pelo agressor para atingir uma Vítima.
  - Massivas mensagens de intimidação foram recebidas em suas redes sociais após sua última fala.
- Maneira [Manner] A Maneira como o agressor age sobre a Vítima.
  - semantic_type: @manner
- Tempo [Time] Identifica o Tempo no qual o evento danoso ocorre.
  - semantic_type: @time

RELAÇÕES:

- Herda de: –
- É Herdado de: –
- Perspectiva sobre: –
- É Perspectivizado em: –
- Usa: Causar_dano
- É utilizado em: –

- Subframe de: Violência
- Têm Subframe(s): –
- Precede: –
- É Precedido por: –
- É Incoativo de: –
- É Causativo de: –
- Ver também: –

APPENDIX C — DIRETRIZES PARA ANOTAÇÃO DE DADOS

Você está sendo convidada/o a ler as diretrizes para identificação de discursos intolerantes

C.1 Discurso Intolerante

Definição Adotada

Discurso intolerante e/ou de ódio é uma linguagem que ataca ou diminui, que incita à violência ou ao ódio contra grupos, com base em características específicas, como aparência física, religião, descendência, nacionalidade ou origem étnica, orientação sexual, identidade de gênero ou outros. Pode ocorrer com diferentes estilos linguísticos, mesmo em formas sutis ou quando o humor é usado. [Fonte: A Survey on Automatic Detection of Hate Speech in Text].

C.2 Características

Considere as seguintes características[1] ao avaliar um discurso enquanto intolerante:

Oposição: Nós x Eles (Temas e figuras de oposição)

* em geral essa categoria se sobrepõe às demais.

Quando há oposição entre um grupo em que as pessoas se identificam por ideias em comum e outro grupo considerado desviante dessas ideias. Em geral, são atribuídos traços frequentes a esses grupos considerados desviantes: traços físicos e características comportamentais de animais; o da "anormalidade" do diferente, que é e age contra a "natureza"; o do caráter doentio e esteticamente condenável da diferença, tais como alguém considerado louco, ou esteticamente feio; o da imoralidade do "outro", de sua falta de ética.

Os falantes geralmente descrevem um grupo externo como biologicamente subumano: como animais, insetos ou mesmo microorganismos como bactérias ou vírus. Persistentemente, em casos de genocídio e atrocidade em massa, apoiadores e perpetradores

[1] As três primeiras características apresentadas derivam do trabalho de Barros (2014), enquanto que a última é uma fruto de estudos feitos para o desenvolvimento deste trabalho.

referem-se às suas vítimas como vermes (ratos, baratas, raposas ou cobras), bestas (macacos ou babuínos) ou perigos biológicos (um vírus, tumores ou uma infecção). Nem toda linguagem que compara pessoas a animais ou outras criaturas não humanas é desumanizante ou perigosa, é claro - é possível comparar uma pessoa a um animal de uma forma que não reduza as barreiras sociais à violência.

Sanção aos maus cumpridores de contratos sociais

O discurso intolerante é, do ponto de vista narrativo, um discurso de sanção aos sujeitos considerados como maus cumpridores de certos contratos sociais (de branqueamento da sociedade, de pureza da língua, de heterossexualidade e outros). E que, portanto, são reconhecidos como maus atores sociais, maus cidadãos (considerados como pretos ignorantes, maus usuários da língua, índios bárbaros, judeus perigosos, árabes fanáticos, homossexuais promíscuos) e, portanto, podendo ser punidos com a perda de direitos, de emprego ou até mesmo com a morte.

Ódio passional e aversão aos diferentes

Predominam, nesses discursos, dois tipos de paixões: as ditas malevolentes (antipatia, ódio, raiva, xenofobia, etc.) e o medo do "diferente" acima mencionados, e dos danos que ele pode causar. O "diferente", ao "mau" usuário da língua, sujeito do ódio em relação ao estrangeiro, aos de outra "cor", orientação sexual ou religião, é também o sujeito do amor à pátria, à sua língua, ao seu grupo étnico, aos de sua cor, à sua religião, ou seja, complementam-se as relações do ódio em relação ao "diferente" e as paixões benevolentes do paixões malevolentes amor aos "iguais".

Falácia com intenção de propagar ódio

Falácias são construídas por raciocínios aparentemente corretos que levam a falsas conclusões. Ao se tratar de discursos intolerantes, não são incomuns situações onde o emissário do discurso faz uso de argumentos que fogem ao contexto em questão, passando então a atacar a pessoa, e não as ideias apontadas por ela, a quem o discurso se dirige, até distorcendo o argumento utilizado por esta.

APPENDIX D — TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Você está sendo convidado(a) a participar como voluntário(a) do estudo **"IDENTIFICAÇÃO E REPRESENTAÇÃO DE DISCURSOS DE ÓDIO EM REDES SOCIAIS"**. A pesquisa tem como objetivo avaliar o uso de indicadores linguísticos característicos a discursos de ódio associado a métodos computacionais como a classificação de textos por algoritmos de aprendizado de máquina (i.e., procedimentos treinados para aprender a partir dos dados recebidos) treinados para isto. Para este propósito, como forma de viabilização de enquadramento de conteúdos disseminados em redes sociais considerando um modelo de representação de nível conceitual (frames), considerando um modelo de enquadramento (i.e., frame) de violência simbólica (i.e., representação conceitual de atos violentos realizados de formas que não necessariamente envolvem violência física) para tal, como uma forma de propor um meio de interposição de discursos intolerantes. Assim, pretendemos compreender a percepção de usuários reais quanto a discursos disseminados em tais meios. É para esta etapa de avaliação que gostaríamos da sua colaboração.

Sua participação consistirá em:

1. Ler e interpretar características propostas (heurísticas linguísticas) como meio de caracterização de discursos de ódio apresentadas na introdução do formulário;

2. Realizar a leitura do frame de violência simbólica apresentado;

3. Avaliar, em uma escala de 0 a 5 com descrições verbais, o grau de compatibilidade dos 10 exemplos apresentados na sequência com a identificação de tais características de discursos de ódio previamente apontadas;

4. Para os mesmos exemplos de texto apresentados, verificar a adequação do proposto para enquadramento de tais discursos segundo uma formalização de acordo com heurísticas linguísticas apontadas na pesquisa.

O tempo estimado de resposta ao questionário é de aproximadamente 20 minutos.

SIGILO E PRIVACIDADE

Os dados pessoais que fornecerá serão utilizados somente para a validação dos critérios de inclusão da pesquisa, e no caso do e-mail, também para envio de uma via

deste documento. Sua identidade será codificada e pesquisadores da equipe terão acesso a esta chave de identificação. Suas respostas e dados pessoais serão guardados em confidencialidade, seu nome não aparecerá durante a pesquisa, nem quando os resultados forem apresentados, e seus dados de contato não serão repassados a outrem.

DESCONFORTO E RISCOS

Responder ao questionário não está associado a riscos de saúde diretos. Entretanto, discursos dispostos nas questões podem causar desconforto ao responder o questionário. É importante ressaltar que a sua participação é voluntária e que você tem todo o direito de suspender e interromper a sua participação a qualquer momento. Você não terá custos em participar desta pesquisa, e também não receberá nenhum benefício financeiro ao fazê-lo.

BENEFÍCIOS

Este estudo não apresenta benefícios diretos aos participantes. Porém, seus resultados poderão ser benéficos para nortear e ampliar estudos voltados à detecção de de discurso de ódio em redes sociais. De forma geral, os benefícios proporcionados por esta pesquisa concentram-se na contribuição para o desenvolvimento do estudo acadêmico centrado em mídias sociais com foco em discursos potencialmente danosos; e ainda a meditação por parte do usuário sobre comportamentos apresentados em tais meios e suas repercussões no cotidiano.

CONTATO

Em caso de dúvidas acerca dos objetivos da pesquisa e/ou dos métodos utilizados, pode entrar em contato com o pesquisador responsável, professor Dr. Leandro Krug Wives, a ser contato através do e-mail: leandro.wives@ufrgs.br; ou ainda entrar em contato com Brenda Salenave Santana, através do e-mail: bssantana@inf.ufrgs.br. Colocamo-nos à disposição para responder a quaisquer dúvidas.

Este documento visa assegurar seus direitos como participante. Ao clicar em "aceito participar desta pesquisa", você declara que concorda em participar da pesquisa e que todas as suas dúvidas foram esclarecidas. A partir de então, você passará a responder o questionário. Ao finalizar as respostas deste questionário, uma via deste termo será encaminhada ao seu e-mail. Desde já agradecemos a sua participação.

Comitê de Ética em Pesquisa/UFRGS

Prédio da Reitoria – 2º andar – Campus Central Av. Paulo Gama, 110 – 90040-060 – Porto Alegre, RS
Telefone: (51) 3308- 3738
E-mail: etica@propesq.ufrgs.br

APPENDIX E — QUESTIONÁRIO

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Você está sendo convidado(a) a participar como voluntário(a) do estudo "**IDENTIFICAÇÃO E REPRESENTAÇÃO DE DISCURSOS DE ÓDIO EM REDES SOCIAIS**". A pesquisa tem como objetivo avaliar o uso de indicadores linguísticos característicos a discursos de ódio associado a métodos computacionais como a classificação de textos por algoritmos de aprendizado de máquina (i.e., procedimentos treinados para aprender a partir dos dados recebidos) treinados para isto. Para este propósito, como forma de viabilização de enquadramento de conteúdos disseminados em redes sociais considerando um modelo de representação de nível conceitual (frames), considerando um modelo de enquadramento (i.e., frame) de violência simbólica (i.e., representação conceitual de atos violentos realizados de formas que não necessariamente envolvem violência física) para tal, como uma forma de propor um meio de interposição de discursos intolerantes. Assim, pretendemos compreender a percepção de usuários reais quanto a discursos disseminados em tais meios. É para esta etapa de avaliação que gostaríamos da sua colaboração.

Sua participação consistirá em:

1. Ler e interpretar características propostas (heurísticas linguísticas) como meio de caracterização de discursos de ódio apresentadas na introdução do formulário;
2. Realizar a leitura do frame de violência simbólica apresentado;
3. Avaliar, em uma escala de 0 a 7 com descrições verbais, o grau de compatibilidade dos 10 exemplos apresentados na sequência com a identificação de tais características de discursos de ódio previamente apontadas;
4. Para os mesmos exemplos de texto apresentados, verificar a adequação do proposto para enquadramento de tais discursos segundo uma formalização de acordo heurísticas linguísticas apontadas na pesquisa.

- O tempo estimado de resposta ao questionário é de aproximadamente 20 minutos.

SIGILO E PRIVACIDADE

Os dados pessoais que fornecerá serão utilizados somente para a validação dos critérios de inclusão da pesquisa, e no caso do e-mail, também para envio de uma via deste documento. Sua identidade será codificada e pesquisadores da equipe terão acesso a esta chave de identificação.
Suas respostas e dados pessoais serão guardados em confidencialidade, seu nome não aparecerá durante a pesquisa, nem quando os resultados forem apresentados, e seus dados de contato não serão repassados a outrem.

DESCONFORTO E RISCOS

Responder ao questionário não está associado a riscos de saúde diretos. Entretanto, discursos dispostos nas questões podem causar desconforto ao responder o questionário. É importante ressaltar que a sua participação é voluntária e que você tem todo o direito de suspender e interromper a sua participação a qualquer momento. Você não terá custos em participar desta pesquisa, e também não receberá nenhum benefício financeiro ao fazê-lo.

BENEFÍCIOS

Este estudo não apresenta benefícios diretos aos participantes. Porém, seus resultados poderão ser benéficos para nortear e ampliar estudos voltados à detecção de de discurso de ódio em redes sociais. De forma geral, os benefícios proporcionados por esta pesquisa concentram-se na contribuição para o desenvolvimento do estudo acadêmico centrado em mídias sociais com foco em discursos potencialmente danosos; e ainda a meditação por parte do usuário sobre comportamentos apresentados em tais meios e suas repercussões no cotidiano.

CONTATO

Em caso de dúvidas acerca dos objetivos da pesquisa e/ou dos métodos utilizados, pode entrar em contato com o pesquisador responsável, professor Dr. Leandro Krug Wives, a ser contato através do e-mail: leandro.wives@ufrgs.br; ou ainda entrar em contato com Brenda Salenave Santana, através do e-mail: bssantana@inf.ufrgs.br. Colocamo-nos à disposição para responder a quaisquer dúvidas.

Este documento visa assegurar seus direitos como participante. Ao clicar em “aceito participar desta pesquisa”, você declara que concorda em participar da pesquisa e que todas as suas dúvidas foram esclarecidas. A partir de então, você passará a responder o questionário. Ao finalizar as respostas deste questionário, uma via deste termo será encaminhada ao seu e-mail. Desde já agradecemos a sua participação.

Comitê de Ética em Pesquisa/UFRGS
Prédio da Reitoria – 2o andar – Campus Central Av. Paulo Gama, 110 – 90040-060 -- Porto Alegre, RS
Telefone: (51) 3308- 3738
E-mail: etica@propesq.ufrgs.br

Continuar

Em caso de dúvidas acerca dos objetivos da pesquisa e/ou dos métodos utilizados, pode entrar em contato com o pesquisador responsável, professor Dr. Leandro Krug Wives (mailto:leandro.wives@ufrgs.br); ou ainda entrar em contato com Brenda Salenave Santana (mailto:bssantana@inf.ufrgs.br).

Antes de começar, nesta etapa, faremos perguntas para conhecer você melhor. Todos os dados serão mantidos em sigilo.

E-mail

Gênero

Idade

Formação

Raça / Etnia

Continuar

Em caso de dúvidas acerca dos objetivos da pesquisa e/ou dos métodos utilizados, pode entrar em contato com o pesquisador responsável, professor Dr. Leandro Krug Wives (mailto:leandro.wives@ufrgs.br); ou ainda entrar em contato com Brenda Salenave Santana (mailto:bssantana@inf.ufrgs.br).

Antes de começar: O que entendemos como Discurso Intolerante

Definição Adotada

Discurso intolerante e/ou de ódio é uma linguagem que ataca ou diminui, que incita à violência ou ao ódio contra grupos, com base em características específicas, como aparência física, religião, descendência, nacionalidade ou origem étnica, orientação sexual, identidade de gênero ou outros. Pode ocorrer com diferentes estilos linguísticos, mesmo em formas sutis ou quando o humor é usado.

Características

1. Oposição: Nós x Eles (Temas e figuras de oposição)

*** em geral essa categoria se sobrepõe às demais**

Quando há oposição entre um grupo em que as pessoas se identificam por ideias em comum e outro grupo considerado desviante dessas ideias. Em geral, são atribuídos traços frequentes a esses grupos considerados desviantes: traços físicos e características comportamentais de animais; o da “anormalidade” do diferente, que é e age contra a “natureza”; o do caráter doentio e esteticamente condenável da diferença, tais como alguém considerado louco, ou esteticamente feio; o da imoralidade do “outro”, de sua falta de ética.

Os falantes geralmente descrevem um grupo externo como biologicamente subumano: como animais, insetos ou mesmo microorganismos como bactérias ou vírus. Persistentemente, em casos de genocídio e atrocidade em massa, apoiadores e perpetradores referem-se às suas vítimas como vermes (ratos, baratas, raposas ou cobras), bestas (macacos ou babuínos) ou perigos biológicos (um vírus, tumores ou uma infecção). Nem toda linguagem que compara pessoas a animais ou outras criaturas não humanas é desumanizante ou perigosa, é claro - é possível comparar uma pessoa a um animal de uma forma que não reduza as barreiras sociais à violência.

2. Sanção aos maus cumpridores de contratos sociais

O discurso intolerante é, do ponto de vista narrativo, um discurso de sanção aos sujeitos considerados como maus cumpridores de certos contratos sociais (de branqueamento da sociedade, de pureza da língua, de heterossexualidade e outros). E que, portanto, são reconhecidos como maus atores sociais, maus cidadãos (considerados como pretos ignorantes, maus usuários da língua, índios bárbaros, judeus perigosos, árabes fanáticos, homossexuais promíscuos) e, portanto, podendo ser punidos com a perda de direitos, de emprego ou até mesmo com a morte.

3. Ódio passional e aversão aos diferentes

Predominam, nesses discursos, dois tipos de paixões: as ditas malevolentes (antipatia, ódio, raiva, xenofobia, etc.) e o medo do "diferente" acima mencionados, e dos danos que ele pode causar. O "diferente", ao "mau" usuário da língua, sujeito do ódio em relação ao ao estrangeiro, aos de outra "cor", direção sexual ou religião,

é também o sujeito do amor à pátria, à sua língua, ao seu grupo étnico, aos de sua cor, à sua religião, ou seja, complementam-se as do ódio em relação ao "diferente" e as paixões benevolentes do paixões malevolentes amor aos "iguais".

4. Falácia com intenção de propagar ódio

Falácias são construídas por raciocínios aparentemente corretos que levam a falsas conclusões. Ao se tratar de discursos intolerantes, não são incomuns situações onde o emissário do discurso faz uso de argumentos que fogem ao contexto em questão, passando então a atacar a pessoa, e não as ideias apontadas por ela, a quem o discurso se dirige, até distorcendo o argumento utilizado por esta."

Continuar

Em caso de dúvidas acerca dos objetivos da pesquisa e/ou dos métodos utilizados, pode entrar em contato com o pesquisador responsável, professor Dr. Leandro Krug Wives (mailto:leandro.wives@ufrgs.br); ou ainda entrar em contato com Brenda Salenave Santana (mailto:bssantana@inf.ufrgs.br).

Você será apresentada/o a um conjunto de 10 tweets de mulheres políticas durante o período eleitoral de 2020, e as respostas (replies) recebidas por elas. Pedimos que, observando **com atenção a essas respostas (replies) dadas a tais tweets**, você avalie se de acordo com as características apresentadas, é possível classificar essas respostas (replies) como discurso intolerante ou não. Ainda, em caso afirmativo, pedimos que, se possível, indique também quais das características se aplicam para esta classificação.

Um guia para as definições das características aqui utilizadas pode ser acessado a qualquer momento em: https://bit.ly/3Ba4kCl (https://bit.ly/3Ba4kCl)

Tweet: Fui assediada publicamente pelo deputado Fernando Cury PS em meio a votacao do orcamento do Estado na ALESP na noite de ontem 16 durante a 65a Sessao Plenaria Extraordinaria da casa

Reply: Senhora Deputada nao concordo com seu partido e a maior parte de suas ideias Porem Assedio e inadmissivel esse Deputado deve ser punido com todo rigor

Contém discurso intolerante?

Não contém

1
2
3
4
5

Contém

Em sua opinião, quais características se aplicam? Você pode escolher mais de uma alternativa.

Frame Violência Simbólica

Considere a seguinte descrição: **"Este frame descreve atos (ou situações caracterizadas por atos) que resultam em uma forma de violência sem dano físico. A violência simbólica causa danos morais, psicológicos, e emocionais. Esses atos simbólicos podem envolver um Agressor (ou Agressores) causando danos simbólicos a uma vítima, ou ainda Agressores causando danos simbólicos uns aos aos outros."**

Caso você tenha considerado a resposta de alguma forma intolerante, essa descrição é suficiente para definir violência simbólica?

Inadequada

Adequada

Considerações adicionais:

Continuar

Em caso de dúvidas acerca dos objetivos da pesquisa e/ou dos métodos utilizados, pode entrar em contato com o pesquisador responsável, professor Dr. Leandro Krug Wives (mailto:leandro.wives@ufrgs.br); ou ainda entrar em contato com Brenda Salenave Santana (mailto:bssantana@inf.ufrgs.br).

Obrigada por responder a pesquisa **"IDENTIFICAÇÃO E REPRESENTAÇÃO DE DISCURSOS DE ÓDIO EM REDES SOCIAIS"**!

Logo uma cópia do Termo de Consentimento Livre e Esclarecido (TCLE) será enviada ao seu e-mail.

Em caso de dúvidas acerca dos objetivos da pesquisa e/ou dos métodos utilizados, pode entrar em contato com o pesquisador responsável, professor Dr. Leandro Krug Wives (mailto:leandro.wives@ufrgs.br); ou ainda entrar em contato com Brenda Salenave Santana (mailto:bssantana@inf.ufrgs.br).

APPENDIX F — ACTIVITIES PERFORMED

In addition to the knowledge and intellectual skills required for the proposal of this thesis, the doctoral period has given support to the academic and pedagogical development of the author, allowing the preparation and professional improvement for her scientific career. In addition to the main activities for developing this research, the author was also involved in different side projects to broaden this research's horizons. Within this, it stands out participation in a discussion group and studies focused on toxic discourses disseminated during a covid-19 pandemic. Among other activities involved in this doctoral process, the dissemination of results achieved through publications (subsection F.1), participation in evaluation boards, program committees and scientific events (subsection F.2), teaching experience (subsection F.3), and the other cooperations carried out (subsection F.4) are described next.

F.1 Publications

The following are the research artifacts already published, submitted for publication, or in the writing process, up to the present moment of the doctorate, in chronological order, i.e., from the oldest to the most current.

1. **Detecting Group Beliefs Related to 2018's Brazilian Elections in Tweets: A Combined Study on Modeling Topics and Sentiment Analysis**

Abstract. 2018's Brazilian presidential elections highlighted the influence of alternative media and social networks, such as Twitter. In this work, we perform an analysis covering politically motivated discourses related to the second round in Brazilian elections. In order to verify whether similar discourses reinforce group engagement to personal beliefs, we collected a set of tweets related to political hashtags at that moment. To this end, we have used a combination of topic modeling approach with opinion mining techniques to analyze the motivated political discourses. Using SentiLex-PT, a Portuguese sentiment lexicon, we extracted from the dataset the top 5 most frequent group of words related to opinions. Applying a bag-of-words model, the cosine similarity calculation was performed between each opinion and the observed groups. This study allowed us to observe an exacerbated use of passionate discourses

in the digital political scenario as a form of appreciation and engagement to the groups which convey similar beliefs.

Status. Published in the Workshop on Digital Humanities and Natural Language Processing 2020; Santana and Vanin (2020)

2. **Report on the third international workshop on narrative extraction from texts (Text2Story 2020)**

Abstract. The Third International Workshop on Narrative Extraction from Texts (Text2Story'20 [<https://text2story20.inesctec.pt/>]) was held on the 14th of April 2020, in conjunction with the 42nd European Conference on Information Retrieval (ECIR 2020). This year due to the Covid-19 outbreak the Text2Story workshop was held online on Zoom platform. During the course of the day, an average of more than 60 attendees had the opportunity to follow-up and discuss the recent advances in extraction and formal representation of narratives. The workshop consisted of two invited keynotes and thirteen paper presentations. The proceedings of the workshop are available online at <http://ceur-ws.org/Vol-2593/>.

Status. Puplished in ACM SIGIR Forum; Campos et al. (2021)

3. **TLS-Covid19: A New Annotated Corpus for Timeline Summarization**

Abstact. The rise of social media and the explosion of digital news in the web sphere have created new challenges to extract knowledge and make sense of published information. Automated timeline generation appears in this context as a promising answer to help users dealing with this information overload problem. Formally, Timeline Summarization (TLS) can be defined as a subtask of Multi-Document Summarization (MDS) conceived to highlight the most important information during the development of a story over time by summarizing long-lasting events in a timely ordered fashion. As opposed to traditional MDS, TLS has a limited number of publicly available datasets. In this paper, we propose TLS-Covid19 dataset, a novel corpus for the Portuguese and English languages. Our aim is to provide a new, larger and multi-lingual TLS annotated dataset that could foster timeline summarization evaluation research and, at the same time, enable the study of news coverage about the COVID-19

pandemic. TLS-Covid19 consists of 178 curated topics related to the COVID-19 outbreak, with associated news articles covering almost the entire year of 2020 and their respective reference timelines as gold-standard. As a final outcome, we conduct an experimental study on the pro-posed dataset over two extreme baseline TLS methods. All the resources are publicly available at <https://github.com/LIAAD/tls-covid19>.

Status. Published on the main track at the 43rd European Conference on Information Retrieval (ECIR) – Qualis A2; Pasquali et al. (2021).

4. **Brat2Viz: a Tool and Pipeline for Visualizing Narratives from Annotated Texts**

Abstract. Narrative Extraction from text is a complex task that starts by identifying a set of narrative elements (actors, events, times), and the semantic links between them (temporal, referential, semantic roles). The outcome is a structure or set of structures which can then be represented graphically, thus opening room for further and alternative exploration of the plot. Such visualization can also be useful during the on-going annotation process. Manual annotation of narratives can be a complex effort and the possibility offered by the Brat annotation tool of annotating directly on the text does not seem sufficiently helpful. In this paper, we propose Brat2Viz, a tool and a pipeline that displays visualization of narrative information annotated in Brat. Brat2Viz reads the annotation file of Brat, produces an intermediate representation in the declarative language DRS (Discourse Representation Structure), and from this obtains the visualization. Currently, we make available two visualization schemes: MSC (Message Sequence Chart) and Knowledge Graphs. The modularity of the pipeline enables the future extension to new annotation sources, different annotation schemes, and alternative visualizations or representations. We illustrate the pipeline using examples from an European Portuguese news corpus.

Status. Published on the Fourth International Workshop on Narrative Extraction from Texts held in conjunction with the 43rd European Conference on Information Retrieval; Amorim et al. (2021).

5. **Sexist Hate Speech: Identifying Potential Online Verbal Violence Instances**

Abstract. Online communication provides space for content dissemination and opinion

sharing. However, the limit between opinion and offense might be exceeded, characterizing hate speech. Moreover, its automatic detection is challenging, and approaches focused on the Portuguese language are scarce. This paper proposes an interface between linguistic concepts and computational interventions to support hate speech detection. We applied a Natural Language Processing pipeline involving topic modeling and semantic role labeling, allowing a semi-automatic identification of hate speech. We also discuss how such speech qualifies as a type of verbal violence widespread on social networks to reinforce a sexist stereotype. Finally, we use Twitter data to analyze information that resulted in virtual attacks against a specific person. As an achievement, this work validates the use of linguistic features to annotate data either as hate speech or not. It also proposes using fallacies as a potential additional feature to identify potential intolerant discourses.

Status. Published in the International Conference on Computational Processing of the Portuguese Language (PROPOR'22); Santana, Vanin and Wives (2022)

6. **A Survey on Narrative Extraction from Textual Data**

Abstract. Narratives are present in many forms of human expression and can be understood as a fundamental way of communication between people. Computational understanding of the underlying story of a narrative, however, may be a rather complex task for both linguists and computational linguistics. Such task can be approached using natural language processing techniques to automatically extract narratives from texts. In this paper, we present an in depth survey of narrative extraction from text, providing a establishing a basis/framework for the study roadmap to the study of this area as a whole as a means to consolidate a view on this line of research. We aim to fulfill the current gap by identifying important research efforts at the crossroad between linguists and computer scientists. In particular, we highlight the importance and complexity of the annotation process, as a crucial step for the training stage. Next, we detail methods and approaches regarding the identification and extraction of narrative components, their linkage and understanding of likely inherent relationships, before detailing formal narrative representation structures as an intermediate step for visualization and data exploration purposes. We then move into the narrative evaluation task aspects, and conclude this survey by

highlighting important open issues under the domain of narratives extraction from texts that are yet to be explored.

Status. Published in Artificial Intelligence Review; Santana et al. (2023).

F.2 Participations

In the quest to improve scientific skills in terms of engagement and collaboration with the academic community, the author was a member of the evaluation committee of specialization and graduation boards (F.2.1). Also, I participated in conferences (F.2.2), served on program committees, and served as primary reviewer and secondary reviewer at different events related to computer studies (F.2.3), including internationally recognized conferences.

F.2.1 Evaluation boards

During the doctoral period, I had the opportunity to join the evaluation committee of eleven *lato sensu* graduate students of the specialization course of *Especialização em Informática Instrumental para Professores da Educação Básica*. In addition to these, I participated as an evaluating board of two bachelors in Biomedical Informatics whose subjects were related to natural language processing and/or data science. The evaluated works are listed below:

- Participation in conclusion work boards in the UFRGS *Especialização em Informática Instrumental para Professores da Educação Básica* course - Specialization level. The titles of the works are listed in the sequence:

    1. *Do uso das Tecnologias da Informação e Comunicação em Sala de Aula, ao Vilão Cyberbullying*
    2. *A Utilização dos Jogos Digitais no Processo de Alfabetização*
    3. *O Plágio em Trabalhos de Pesquisa dos Alunos do Ensino Fundamental e Médio: Um Estudo de Caso*
    4. *Utilização do Jogo Geoguessr para Ensino de Geografia em uma Escola Pública de Ensino Fundamental do Município de Sapucaia do Sul - RS*

5. *Educação à Distância: Propriedades, Estratégias e Competências da Tutoria*
6. *Alfabetização e Tecnologia: Uma Revisão de Literatura*
7. *Recursos Educacionais Abertos para Todos - Gurias nas Exatas Odila*
8. *Dispositivos Móveis como Instrumento para Inclusão de Deficientes Visuais: Uma visão a partir da Teoria Sócio-Histórica*
9. *Uso De Tecnologia No Atendimento A Alunos Com Altas Habilidades / Superdotação*
10. *Smartphones - Instrumentos de apoio ao ensino de matemática no Ensino Fundamental*
11. *Uso De Aplicativos no Ensino Da Matemática*

- Participation in conclusion work boards in the UFSCPA's Biomedical Informatics course - Bachelor's level. The titles of the works are listed in the sequence:

1. *Seleção de Estruturas Representativas De Proteínas: Um Estudo Sobre Algoritmos de Aprendizado de Máquina em Dados de Simulações de Dinâmica Molecular.* - 2019;
2. *Extração e Recuperação de Informações em Documentos Científicos sobre a COVID-19 Utilizando Processamento de Linguagem Natural* - 2020;

F.2.2 Scientific events

1. Participation in the 4th edition of the International Conference on the Computational Processing of Portuguese (PROPOR 2020); and also attended the tutorial entitled *Evolução dos Modelos de Linguagem* (Evolution of Language Models) given during the event;

2. Participation in the 10th Lisbon Machine Learning School - LxMLS 2020 organized by the Instituto Superior Técnico (IST) of Portugal, to obtain complementary training in the techniques studied for this proposal;

3. Participation in the 5th edition of the International Conference on the Computational Processing of Portuguese (PROPOR 2022);

F.2.3 Program Committees and Reviews

1. Acting as a secondary reviewer in the Empirical Methods in Natural Language Processing (EMNLP) 2020 Conference;

2. Acting as a secondary reviewer in the International Conference on Theory and Practice of Digital Libraries (TPDL) 2020;

3. Acting as a secondary reviewer in the Special Interest Group on Information Retrieval (SIGIR) 2020 conference;

4. Acting as a secondary reviewer in the 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'20);

5. Acting as reviewer in the *Jornadas Iberoamericanas de Ingeniería de Software e Ingeniería del Conocimiento* (JIISIC2020);

6. Acting as a secondary reviewer in the Driving Simulation Conference (DSC) Europe 2020;

7. Acting as a secondary reviewer in the Conference on Computational Natural Language Learning (CoNLL) 2020;

8. Acting as a secondary reviewer in the European Conference on Information Retrieval (ECIR) 2021 conference;

9. Participation of the organizing committee of the Artificial Intelligence for Narratives Workshop (AI4Narratives'21);

10. Acting as a reviewer in the Fourth International Workshop on Narrative Extraction from Texts (Text2Story'21);

    - Recognized Reviewer Award.

11. Acting as a secondary reviewer in the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021);

12. Acting as a secondary reviewer in the 16th Doctoral Symposium in Informatics Engineering (DSIE 2021);

13. Acting as a secondary reviewer for the demo papers in the Special Interest Group on Information Retrieval (SIGIR) 2021 conference;

14. Acting as a secondary reviewer in the International Conference on Theory and Practice of Digital Libraries (TPDL) 2021;

15. Acting as a reviewer in the 44th European Conference on Information Retrieval (ECIR'22), which led to the receipt of an outstanding reviewer award.

    - Outstanding Reviewer Award.

16. Acting as a reviewer for the Expert Systems with Applications Journal.

F.3 Teaching and Co guidances

1. Compliance with Teaching Practice II "*Pesquisa e Classificação de Dados*" at the undergraduate level under the supervision of the Prof. Leandro Krug Wives;

2. Co guidance in a work entitled "*Análise de polaridade de opinião nas redes sociais no setor bancário*" in 2019 - Specialization level;

3. Acting as an assistant professor in the discipline of "*Text Mining Aplicado à Business Analytics*" offered to the course of Business Analytics by the Escola de Administração da UFRGS in 2019 and 2020;

4. Co guidance in a work entitled "*Otimizando o processo de brainstorming com técnicas de Processamento de Linguagem Natural e Aprendizado de Máquina*" in 2021 - Bachelor's level;

F.4 Cooperations

Besides, cooperations were carried out in activities and external projects that contributed to the attribution of knowledge related to research and dissemination of knowledge learned in the academic environment.

F.4.1 Cooperations focused on the business environment

1. Acting as coordinator in the Artificial Intelligence and Machine Learning Track at The Developers Conference (TDC) 2020.

2. Acting on the *Saúde com agente*[1], with the artificial intelligence team responsible for verifying activity reports submitted by parties involved in the project..

F.4.2 International Cooperation

In the first month after entering the doctorate, I was selected to work on a project linked to the Artificial Intelligence and Decision Support (LIAAD) INESC TEC laboratory's and the University of Porto to apply natural language processing methods in Portuguese texts. Due to the necessary bureaucratic procedures, the beginning of this collaboration took place only from January 1, 2020.

I was allocated to the project named Text2Story. This project is motivated by the multiple formats, mainly through the web and specific internet-based applications running on smartphones and tablets, that nowadays journalistic content is distributed in multiple formats. Being text an essential format, but readers (or more accurately, users or information consumers) heavily rely on images, videos, slideshows, charts, and infographics. Textual content is still the primary representation of information. This vibrant research line poses many challenging problems in information extraction and automatic production of media content. In this project, the researchers want to be able to extract narratives/stories from news articles or collections of related news articles (unstructured data) about the same (or related) subject, representing those narratives in intermediate data structures (structured data) and making this available to subsequent media production processes (semi-automatic generation of slide shows, infographics and other visualizations, video sequences, games, etc.). In summary, the Text2Story project aims to develop a conceptual framework and operational pipeline to extract narratives from textual sources. The project focuses on the automatic processing of journalistic text in written Portuguese.

In this project, I worked on the study of NLP practices focused on extracting relationships between temporal data and events in narrative texts and the representation in the declarative language DRS (Discourse Representation Structure) of previously anno-

[1] See <https://saudecomagente.ufrgs.br/saude/>

tated narratives. I was also the webmaster of the project's dissemination page <https://text2story.inesctec.pt>. This collaboration lasted until the end of November 2022.