FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes

Paula Fortuna

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Sérgio Nunes

June 23, 2017

Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes

Paula Fortuna

Mestrado Integrado em Engenharia Informática e Computação

June 23, 2017

Abstract

The use of the internet and social networks, in particular for communication, has significantly increased in recent years. This growth has also resulted in the adoption of more aggressive communication. Therefore it is important that governments and social network platforms have tools to detect this type of communication, because it can be harmful to its targets. In this thesis we investigate the problem of detecting hate speech posted online.

The first goal of our work was to make a complete overview on the topic, focusing on the perspective of computer science and engineering. We adopted an exhaustive and methodical approach that we called Systematic Literature Review. As a result, we critically summarized different perspectives on the hate speech concept and complemented our definition with rules, examples, and a comparison with other related concepts, such as cyberbullying and abusive language. Regarding the past work in the topic, we observed that the majority of the studies tackles this problem as a machine learning classification task and the studies use either general text mining features (e.g. n-grams, word2vec), or hate speech specific features (e.g. othering discourse). In the majority of these studies new datasets are collected, but those remain private, which makes more difficult to compare results across different works. We concluded also that this field is still in an early stage, with several open research opportunities.

As we found no research on the topic in Portuguese, the second goal of this work was to annotate a dataset for this language and to make it available as well. Regarding the dataset annotation, we built a classification system using a hierarchical structure. The main advantage of this strategy is that it allows to better consider nuances in the hate speech concept, such as the existence and intersectionality of the subtypes of hate speech. Our data was collected from Twitter, and manually annotated by following a set of rules, that are also a valuable product of our work.

We annotated a dataset with 5,668 messages from 1,156 distinct users, where 85 distinct classes of hate speech were considered. From the total 5,668 messages, around 22% contain some type of hate speech. Regarding the annotators agreement, using the hierarchical approach allowed us to improve results, however this was still an issue in identifying hate speech. Further analysis pointed out that the several types of hate speech present different characteristics (e.g. distinct number of messages, time occurrences, vocabulary size, distinct n-grams and POS).

A final goal of our thesis was to investigate the potential advantages of using hierarchical classes to annotate a dataset. For this, we used the dataset annotated for Portuguese and we conducted an experiment with training, validation and test phases. In this experiment we compare two different approaches: we called unimodel to the model using only the hate speech class; and multimodel to the model using the several hierarchical classes. The main conclusion of our experiment was that the performance of the multimodel seemed to be slightly better than the unimodel in the F1 metric, and additionally, our method helped to identify a larger number of hate speech messages. This is the case because it has a better recall, in detriment of the precision.

Finally, we think that in the future this experiment can be extended in order to better identify hate speech and the respective subtypes.

Resumo

A internet e as redes sociais têm vindo a ser cada vez mais utilizadas como ferramentas de comunicação. Esta utilização, resultou também no crescimento da comunicação agressiva, e uma vez que esta pode ter consequências negativas para os seus alvos, é importante que tanto as autoridades como as plataformas que gerem redes sociais tenham forma de a detetar. No âmbito desta tese investiga-se em particular a deteção automática de discurso de ódio em texto publicado online.

Em primeiro, conduziu-se uma revisão de literatura sobre o tópico, focando-se principalmente na perpetiva da ciência dos computadores e engenharia. Foi adoptado um método sitemático e exaustivo para recolha de documentos. Como resultado, obteve-se uma definição sumária de discurso de ódio, que foi complementada com regras, exemplos e comparada com outros conceitos relacionados (e.g. *Cyberbyllying* e *Abusive language*). Relativamente aos trabalhos na área, observou-se que a maioria considera o problema como uma tarefa de classificação, utilizando aprendizagem automática. Os estudos, usam tanto caraterísticas comuns de *text mining* (e.g. *n-grams*, *word2vec*), como também específicas da deteção de discurso de ódio (e.g. *othering discourse*). Para além disso, na maioria dos estudos os dados utilizados não são disponibilizados posterioremente, o que dificulta a comparação de resultados entre as várias abordagens. Concluiu-se ainda que esta área se encontra ainda numa fase preliminar, com diversas questões de investigação em aberto.

Na revisão de literatura realizada, não foram encontrados estudos em português. Assim, o segundo objetivo deste trabalho foi anotar um dataset para esta língua. Para isso utilizou-se uma estrutura hierárquica de classes, o que tem como vantagem permitir que melhor se considerem as nuances do conceito de discurso de ódio, tais como a interseção dos vários subtipos. As mensagens foram recolhidas no Twitter e manualmente anotadas, seguindo um conjunto de regras, que é também um produto deste trabalho de investigação. O *dataset* recolhido contém 5668 mensagens de 1156 utilizadores diferentes. Foram consideradas 85 subclasses de discurso de ódio, e cerca de 22% das mensagens revelou conter pelo menos uma destas classes. Relativamente ao acordo entre anotadores, usando a abordagem hierárquica foi possível melhorar resultados. Contudo, o acordo continua a ser difícil na identificação de discurso de ódio. Uma análise do conjunto de dados permitiu ainda concluir que os subtipos de discurso de ódio apresentam caraterísticas diferentes (e.g. diferentes número de mensagens, distribuição no tempo, vocabulário, *n-grams* e POS *tags*).

Por fim, investigaram-se as vantagens de se utilizar uma classificação hierárquica na deteção de discurso de ódio. Para isso, conduziu-se uma experiência (com fase de treino, validação e teste) com o conjunto de dados anotado em português. Nesta experiência compararam-se duas abordagens diferentes: *unimodel*, onde o modelo considera apenas a classe discurso de ódio; e *multimodel*, onde o modelo considera diversas classes organizadas hierarquicamente. Concluiu-se que a performance do *multimodel* (F1) parece supererar a do *unimodel*. Adicionalmente, o *multimodel* permite identificar um maior número de mensagens com discurso de ódio, uma vez que tem melhores valores da taxa de recuperação, em detrimento da precisão.

Finalmente, no futuro esta experiência poderá ser extendida, para uma melhor identificação automática de discurso de ódio e respetivos subtipos.

Acknowledgements

I would like to express my gratitude to all the people that supported my research in this topic...

To my family for the love!

To the friends for the adventures and good caring words!

To the Infolab colleagues that always provided me an excellent work environment!

To prof. Sérgio for the trust and supervision.

To all the researchers that make their work freely available to others.

To all the people that have the courage to don't accept all the norms, and keep dreaming and fighting against oppression.

And finally, to all activists that write pedagogical and inspirational texts!

Paula Fortuna

Contents

List of Figures

LIST OF FIGURES

List of Tables

Abbreviations

AAAI	Association for the Advancement of Artificial Intelligence
ACM	Association for Computing Machinery
Acc	Accuracy
AJC	American Jewish Committee
API	Application programming interface
AUC	Area Under Curve
BOW	Bag Of Words
DAG	Directed Acyclic Graph
DAGSVM	Directed Acyclic Graph-Support Vector Machine
DFTPN	Difference between Frequencies of Token in the Positive and Negative messages from the class
EU	European Union
F	F-measure
FBI	Federal Bureau of Investigation
FN	False Negatives
FP	False Positives
HQEH	Homem que é homem
HTML	Hypertext Markup Language
ILGA	International Lesbian and Gay Association
KDIR	Knowledge Discovery and Information Retrieval
LGBTI	Lesbian, gay, bisexual, transgender and intersex
LogReg	Jogistic Regression
MLP	Multilayer Perceptron
NER	Named-entity recognition
NLP	Natural Language Processing
P	Precision
PNR	Partido Nacional Renovador
POS	Part-of-speech
R	Recall
RF	Random Forest
RFTSM	Relative Frequency of a Token in a Set of Messages
Rpart	Recursive partitioning
RT	Retweet
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
TF-IDF	Term frequency–inverse document frequency
TN	True Negatives
TP	True Positives
URL	Uniform Resource Locator
USA	United States of America
Xgboost	Extreme Gradient Boosting

Chapter 1

Introduction

Nowadays people are using more and more social networks to communicate their opinions, share information and search for experiences. The internet allows people to protect themselves behind the screen and interact with each other in a more anonymous environment. Therefore, in social networks, people can have more easily the feeling of being de-individualized and can incur more in aggressive communication [BW15]. In this context, it is important that governments and social network platforms have tools to detect aggressive behavior in general, and hate speech in particular, because it can have negative consequences to its targets. In the scope of this thesis we have several goals. Generally, we aim to enrich this field of research and contribute for solutions to the problem of automatic detection of hate speech online. We analyse the goals of our work more in detail.

1.1 Goals of the work

In the scope of this work we have three different main goals. The first goal of this thesis is to understand what has been done so far in this field. There is little previous literature on identifying hate speech [WH12a] and describing the state of the art in this area is not simple. First, hate speech detection in text is a sub-area within text mining, that intersects with other sub areas (e.g. sentiment analysis). Besides, it is important to distinguish hate speech from other concepts also studied (e.g. cyberbullying). We should also consider that hate speech is object of analysis in other different areas of knowledge, as social sciences and law. However, when we focus on a computer science and engineering point of view, the number of studies in the area is more limited. With the purpose of having an overview of the area and with all these challenges in mind, our review is distinct: we intend it to be exhaustive, methodical and also useful for researchers starting in the topic. We conducted a Systematic Literature Review.

A second goal of our work is to collect a dataset for Portuguese, for the hate speech automatic detection classification task and make it publicly available. This is an important contribution because Portuguese is one of the most spoken languages in the world, and also on social networks [Fox13]. Besides, in our annotation task we try to take into account the nuances of the concept of

hate speech. We consider this classification task as a hierarchical class problem as we present in our third chapter.

A final goal of our thesis is to investigate the potential advantages of using hierarchical classes to annotate a dataset. We use our dataset annotated for Portuguese and we conduct an experiment where we use the information about the subtypes of hate as features in order to understand if these can help in predicting and classifying messages as hate speech.

In order to achieve these goals we divided the thesis in different sections, that we present here briefly.

1.2 Outline of the thesis

Regarding the outline of our work, we dedicate our first chapter to introduce our goals and approach.

In a second chapter we present an overview on the topic of automatic detection of hate speech in text, and we spot some opportunities in this field. One of these opportunities is the lack of research for Portuguese, and therefore we annotate a hate speech dataset for this language, which is presented in the third chapter.

In the subsequent chapter we investigate the advantages of our annotation procedure, that uses a hierarchical structure, and we try to predict hate speech in Portuguese.

In the last chapter we evaluate the achieved goals of our thesis and the future work.

1.3 Language concerns

Finally, it is important to point out that some examples in this thesis use hate speech and also profanity language. These are used just for better understanding of concepts and do not express the point of view of the author.

Chapter 2

Automatic detection of hate speech in text: an overview

This section aims to present the work conducted so far in hate speech automatic detection in text. We provide a systematic approach, that presents not only theoretical aspects, but also practical resources, such as datasets or other projects in the field. We found other articles that already describe key areas explored to automatically detect hate speech using natural language processing (e.g. [SW17]). However, with our systematic research we want to give a complementary analysis, providing more articles, and also practical resources in the field.

In the first subsection "What is Hate Speech?" we analyse the more theoretical aspects of studying this topic: we distinguish amongst different definitions of the concept, analyse particular cases of hate speech, relate hate speech with other close concepts, see how hate speech online has been evolving and who are the main targets of it. After, we also consider "Why to Study Hate Speech Automatic Detection?" which brings attention to the motivation for studying this field.

In order to answer to our third question "What Has Been Done So Far In Automatic Hate Speech Detection Research?", we conducted a Systematic Literature Review, whose method and results are also presented in this section. We try to summarize both quantitative data (e.g. evolution of the number of publications by year, main conferences) and qualitative data (e.g. feature extraction conducted in the field).

In a final part of this section we present the available datasets, projects and conferences. We take into consideration possible difficulties, what to do after detecting hate speech online, and finally we spot some opportunities in the area, as well. We start this section by clarifying concepts, which is a very important task because hate speech is not a clear concept and it is studied in several disciplines.

2.1 What is hate speech?

In this section we present the definitions that are important to clarify in the problem of hate speech automatic detection. We analyse here different perspectives on the hate speech definition and also

other related concepts.

2.1.1 Definitions from several sources

Deciding if a portion of text has hate speech is not linear, even for humans. Thus, if we want machines to identify hate speech, it is important to clearly define this concept in order to make this task easier [RRC+17]. Different sources define hate speech and in this sub-section we try to collect these definitions. We aim to check if there is consensus in the diverse perspectives (Table 2.1 and Table 2.2).

Table 2.1: Hate speech example definitions.

Source	Definition
Code of conduct, between EU and companies [Her16]	"All conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic" [Her16]
ILGA	"Hate speech is public expressions which spread, incite, promote or justify hatred, discrimination or hostility towards a specific group. They contribute to a general climate of intolerance which in turn makes attacks more probable against those given groups." [ILG16]
Scientific paper	"Language which attacks or demeans a group based on race, ethnic origin, religion, disability, gender, age, disability, or sexual orientation/gender identity." [NTT+16]
Facebook	"Content that attacks people based on their actual or perceived race, ethnicity, national origin, religion, sex, gender or gender identity, sexual orientation, disability or disease is not allowed. We do, however, allow clear attempts at humor or satire that might otherwise be considered a possible threat or attack. This includes content that many people may find to be in bad taste (ex: jokes, stand-up comedy, popular song lyrics, etc.)." [Fac13]
Youtube	"We encourage free speech and try to defend your right to express unpopular points of view, but we don't permit hate speech. Hate speech refers to content that promotes violence or hatred against individuals or groups based on certain attributes, such as race or ethnic origin, religion, disability, gender, age, veteran status and sexual orientation/gender identity. There is a fine line between what is and what is not considered to be hate speech. For instance, it is generally okay to criticize a nation-state, but not okay to post malicious hateful comments about a group of people solely based on their ethnicity."[You17]
Twitter	"Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories." [Twi17]

In what concerns to the sources of the definitions (Table 2.1), we collected them from a wide range of origins for several reasons:

- European Union Commission (source Code of conduct on Table 2.1), because they regulate other institutions.
- International minorities associations (source ILGA), because they aim to protect people that are usually targets of hate speech.
- Scientific papers, to include also a perspective from the scientific community.
- Social networks conditions and terms (definitions from Facebook, Youtube and Twitter), because these are some of the main social networks and in these platforms hate speech occurs regularly.

In order to better understand the definitions found, in Table 2.2 we have different columns: the source of the definition (shared with Table 2.1); and four dimensions in which the definitions can be compared ("Hate speech has specific targets", "Hate speech is to incite violence or hate", "Hate speech is to attack or diminish", "Humour has a specific status").

Table 2.2: Hate speech definitions content analysis.

Source	**Hate speech is to incite violence or hate**	**Hate speech is to attack or diminish**	**Hate speech has specific targets**	**Humour has a specific status**
Code of conduct	Yes	Not referenced	Yes	Not referenced
ILGA	Yes	Not referenced	Yes	Not referenced
Scientific paper	Not referenced	Yes	Yes	Not referenced
Facebook	Not referenced	Yes	Yes	Yes
Youtube	Yes	Not referenced	Yes	Not referenced
Twitter	Yes	Yes	Yes	Not referenced

These columns are the result of a manual analysis of the definitions, within a method similar to content analysis [Kri04]. We explain better and analyse these four dimensions in the next paragraphs.

- **Hate speech has specific targets** - All the definitions point out that hate speech has specific targets and it is based on specific characteristics of groups, like ethnic origin, religion or other.
- **Hate speech is to incite violence or hate** - The several definitions use slightly different terms to describe when hate speech occurs. The majority of the definitions point out that hate speech is to incite violence or hate towards a minority (definitions from Code of conduct, ILGA, Youtube, Twitter).
- **Hate speech is to attack or diminish** - Additionally, some other definitions state that hate speech is to use language that attacks or diminishes these groups (definitions from Facebook, Youtube, Twitter).
- **Humour has a specific status** - On the other hand, Facebook points out that some offensive and humorous content is allowed (definition from Facebook). The exceptional status of

> humour makes the boundaries about what is forbidden in the platform more difficult to identify.

After the conclusions we presented previously, we use these four dimensions of analysis to define what is hate speech in the scope of this thesis. We present this definition in the next sub-section.

2.1.2 Our definition of hate speech

Based on the content analysis presented in the previous sub-section, we define hate speech in the scope of this thesis as:

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

To make it even more clear, this definition aims to join the different perspectives presented previously. However, we should also point out that if, on one hand, violence can occur physically and explicitly, on the other hand it can also be subtle. This is the case when stereotypes are reinforced, giving a justification to discrimination and negative bias towards these groups. Therefore, we consider that all subtle forms of discrimination, even jokes, must be marked as hate speech. This is the case because this type of jokes reflect relations between the groups of the jokers and the groups targeted by the jokes, racial relations and stereotypes [KvdE16]. Moreover, repeating these jokes can become a way of reinforcing racist attitudes [Kom16] and, although they are considered harmless, they also have negative psychological effects for some people [DMEY16].

In the next sub-section we clarify the notion of hate speech and analyse some particular cases and examples.

2.1.3 Particular cases and examples of hate speech

In order to better understand how the definition of hate speech can be applied in hate speech automatic detection in text, we should consider some examples and issues. It is important to present a clear definition of the concept to make the task of automatic hate speech detection easier [RRC+17]. One article [KG16] reveals some definitions and cases that Facebook uses to train its workers in the process of handling messages tagged as spam for hate speech. According to Facebook, a message has hate speech when two conditions are met:

- a verbal attack occurs.
- the target of the attack is from a "protected category" (religious affiliation, national origin, etc.).

Some rules extracted from this article are:

- members of religious groups are protected, religion itself is not.
- speaking badly about countries (e.g. France or Germany) is allowed in general, however condemning people on the basis of their nationality is not.
- a protected category combined with another protected category results in yet another protected category (e.g. if someone writes "Irish women are dumb," they would be breaking the rules and their post would be deleted, because "national origins" and "sex" categories apply).
- combining a protected category with an unprotected category, however results in an unprotected category. For this reason, the sentence "Irish teenagers are dumb" does not need to be deleted because the term teenager does not enjoy special protection.
- saying "fucking Muslims" is not allowed, as religious affiliation is a protected category.
- However, the sentence "fucking migrants" is allowed, as migrants are only a "quasi protected category" – a special form that was introduced after complaints were made. This rule states that promoting hate against migrants is allowed under certain circumstances: statements such as "migrants are dirty" are allowed, while "migrants are dirt" is not.

According to this article, some sentences are used to exemplify what should be marked as hate speech (Table 2.3). The examples marked as "violating" should be deleted by the workers, whereas the examples marked as "non-violating" should be ignored.

Table 2.3: Text messages classified by Facebook (Table from [KG16]).

Message	**Evaluation**
Don't trust boys!	violating - delete
Refugees should face the figuring squad!	violating - delete
Fucking Muslims!	violating - delete
Fucking migrants!	non-violating - Ignore
Migrants are filthy cockroaches that will infect our country	violating - delete
I'm such a faggot people call me diva!	non-violating - Ignore
The French are alcoholics	violating - delete
All English people are dirty!	violating - delete
Don't try to explain - Irish Catholics are just idiots	violating - delete
Migrants are scum!	violating - delete
People should stop to use the word nigger.	non-violating - Ignore
I hate migrants!	non-violating - Ignore
Don't trust boys who say they love you!	non-violating - Ignore
Tall girls are just freaks!	non-violating - Ignore
American shitheads!	violating - delete
Migrants are so filthy!	non-violating - Ignore
Refugees! More like rape-fugees!	violating - delete
Asylum seekers out!	violating - delete
Group for blacks only!	non-violating - Ignore

The rules presented by Facebook are arguable. From our point of view there is no reason to restrain hate speech to specific "protected categories". First, because new targets of hate speech

can appear, and in this case these are undetectable unless we redefine these "protected categories". Besides, prejudice can occur even when protected categories are not specifically implied. For instance, boys and men receive at an early age confining and stereotypical messages, that come from family, peers or media, telling them how to behave and feel, relate to each other girls and women. Some of these messages are harmful and have short and long-term consequences for themselves, women, their families, their community and society as a whole [Lea15].

Based on other sources we consider the following rules as a guide for the classification:

- First, usage of disparaging terms and racial epithets with the intent to harm must be considered hate speech.
- However, in a discussion of the words themselves such expressions might be acceptable [WH12b].
- Sometimes these words are used by a speaker who belongs to the targeted group, in order to show pride for belonging to the group. For our purpose, and if there is no contextual clue about it, such terms are categorized as hateful [WH12b].
- Also, references to an organization associated with hate crimes does not by itself constitute hate speech. For instance the name "Ku Klux Klan" is not hateful, as it may appear in historical articles or other legitimate communication [WH12b].
- However, while the endorsement of organizations that promote hate speech does not constitute a verbal attack on another group, in the scope of this work we define that this must be marked as hate speech. In this point, we oppose to the perspective of some other authors [WH12b].
- Besides, calling attention to the fact that an individual belongs to a group and invoking a well known and disparaging stereotype about that group is also hate speech [WH12b].
- Making generalized negative statements about minority groups as in "the refugees will live off our money" is hate speech, due to the incitation of a negative bias towards the group. However, some authors [RRC+17] were unsure about this example as being hate speech.
- We can say that if a text "uses a sexist or racial slur" it contains hate speech [RRC+17].
- However it is also important to point out that the use of some words like "black", "white", "filthy", or other, is marked as hate speech only in some circumstances. Outside of context, these words bear no racial undertones of their own [KW13].
- Speaking badly about countries or religions (e.g. France, Portugal, Catholicism, Islam) is allowed in general, but discrimination is not allowed based on these categories.
- Finally, hate speech can also occur when the statement about the superiority of the in-group are made. This rule was extracted from the Facebook set, discussed in this section.

The presented rules point out that we aim to have a more inclusive and general definition about hate speech than some other perspectives found in literature. This is the case because we want to be able to better describe subtle forms of discrimination amongst the internet and social networks.

2.1.4 Hate speech and other related concepts

Several concepts related with hate speech have been found in the literature: hate [Tar16], cyberbullying [Che11], abusive language [NTT+16], discrimination [Tho16a], profanity [Dic17], toxicity [Jig17] and flaming [GHH07]. In this sub-section we distinguish between these concepts and hate speech (Table 2.4).

Table 2.4: Comparison between hate speech definition and related concepts.

Concept	Definition	Distinction from hate speech
Hate	Expression of hostility without any stated explanation for it [Tar16].	Hate speech is not general hate, it focus towards stereotypes.
Cyberbullying	Aggressive and intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time, against a victim who can not easily defend him or herself [Che11].	Hate speech is more general and not necessarily focused on a specific person.
Discrimination	Process through which a difference is identified and then used as the basis of unfair treatment [Tho16a].	Hate speech is a way of discriminating through verbal means.
Flaming	Flaming are hostile, profane and intimidating comments that can disrupt participation in a community [GHH07]	Hate speech can occur in any content, whereas flaming is aimed towards a participant in the specific context of a discussion.
Abusive language	The term abusive language was used to refer to hurtful language and includes hate speech, derogatory language and also profanity [NTT+16].	Hate speech is a type of abusive language.
Profanity	Offensive or obscene word or phrase [Dic17].	Hate speech can use profanity, but not necessarily.
Toxic language or comment	Toxic is defined as rude, disrespectful or unreasonable comments that are likely to make a person to leave a discussion [Jig17].	Not all toxic comments contain hate speech. Also some hate speech can make people discuss more.

Besides, it is also important to clarify each type of hate speech that we found in literature (Table 2.5).

The concepts presented in this section are distinct from hate speech, however they are related. Therefore, literature and empirical studies focusing on them can give insight about how to automatically detect hate speech.

2.2 Why to study hate speech automatic detection?

Hate speech has become a popular concept over the past few years. There are several reasons to study hate speech automatic detection, which we present in this section.

Table 2.5: Types of hate speech and examples (Table from [SMC+16]).

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

- **European Union Commission directives** - In the last years, different programs are being developed towards the fight against hate speech (e.g. No Hate Speech Movement by the Council of Europe [17]). Recently, the EU Commission pressured Facebook, Youtube, Twitter and Microsoft to sign an EU hate speech code [Her16]. This includes the requirement to review the majority of valid notifications for removal of illegal hate speech in less than 24 hours [Her16, RRC+17].

- **Automatic techniques not available** - automated techniques aim to programmatically classify text as hate speech, making its detection easier for the ones that have the responsibility to protect the public [BW16, RRC+17]. These techniques can be used in order to give a response in less than 24 hours, as demanded for some social networks (citation on this topic in the bullet European Union Commission directives). Some studies have been conducted about hate speech automatic detection, but they did not provide tools.

- **Lack of data about hate speech** - There is a general lack of systematic monitoring, documentation and data collection of hate and violence, namely against LGBTI people [ILG16]. Nevertheless, detecting hate speech is a very important task because it is connected with actual hate crimes [WH16, ILG16, RRC+17] and automatic hate speech detection in text makes it possible to search for it on the internet.

- **Hate Speech Removal** - Some companies and platforms might be interested in hate speech detection and removal [Wen15]. For instance, some newspapers, or even Twitter, need to attract advertisers and therefore cannot risk becoming known as platforms for hate speech [HTB16].

- **Quality of service** [OC14] - Social media companies provide a service: they ease the communication between its users. They profit from this service and therefore they can assume public obligations with respect to the contents transmitted. In this case, quality of service would be: take steps to discourage online hate and remove hate speech within reasonable time. Both can be measured and compared to a standard imposed through legislation.

In addition to all the motivations presented, hate speech automatic detection in text is a topic where some research opportunities exist, as we will see in the next section.

2.3 What has been done so far in automatic hate speech detection research?

In order to describe the state of the art in this field we conducted a Systematic Literature Review. We decided to take a systematic approach because of the lack of summarized information on the topic. In this section, we describe this method and achieved results in detail. We use the term document as a synonym for papers, thesis or any other sort of text document.

2.3.1 Systematic Literature Review

The goal of this Literature Review is to collect the largest possible number of documents in the area of hate speech automatic detection in text. In order to achieve this result we developed a method that we present in the next section.

2.3.1.1 Method description

The method is structured in four phases that are presented and summarized in Figure 2.1.

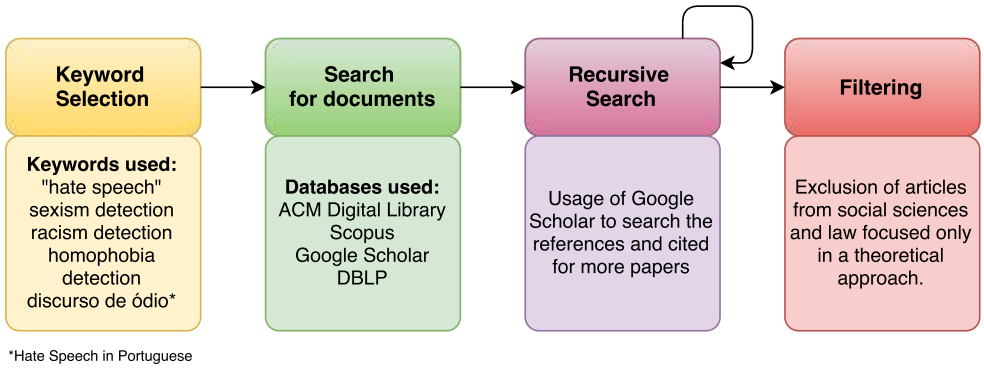

Figure 2.1: Methodology for document collection.

We also present the different phases with more detail in the next paragraphs.

- **Keyword selection** The first phase conducted was the keywords selection. We bear in mind that hate speech is a concept that became more popular recently. Therefore some other related concepts could have been used in the past by the scientific community. We considered terms referring to particular types of hate speech (sexism, racism and homophobia). Besides, we also considered search for “hate speech” in other languages (Portuguese and Spanish).

- **Search for documents** We searched documents in different databases, aiming to gather the largest possible number in the areas of computer science and engineering. Databases from other scientific areas were not considered.

- **Recursive search** We used Google Scholar to get both the references and documents that cite the original work. We check on these two sets and search for the expression "hate speech" on the titles of the candidate documents. We repeated the recursive search with the new documents found.

- **Filtering** An initial step of filtering was conducted. Documents from social sciences and law were immediately deleted.

2.3.1.2 Documents collection and annotation

The process of collecting documents was conducted from September 1st, 2016 to May 18th, 2017. We ended up with a total of 127 documents that we described using the following metrics:

- Name.

- Area of knowledge (we created the categories: "Law and Social Sciences" and "Computer Science and Engineering").

- Conference or journal name.

- Keywords in the document.

- Particular hate (while some articles focus generally in hate speech, others focus in particular types, such as racism).

- Social network (refers to the network used to get samples of text).

- Number of instances used (refers to the size of the dataset used in the work).

- Algorithms used.

- Type of document (we created the categories: "algorithms about hate speech", "algorithms but not about hate speech", "descriptive statistics about hate speech", "descriptives statistics but but not about hate speech" and "theoretical").

- Year of the document.

In the next sections we present the main results from our Systematic Literature Review.

2.3.1.3 Area of knowledge

We classified each document as "Law and Social Sciences" or "Computer Science and Engineering". We concluded that the majority of the works we found is from the first category (N = 76), whereas only few articles are more related with "Computer Science and Engineering" (N = 44). In the scope of this work we are only interested in analyse the papers from the set "Computer Science and Engineering". In the following sections we only focus on this group.

2.3.1.4 Year of the document

As we can see in Figure 2.2, before 2014 the number of documents related with hate speech, from the type "Computer Science and Engineering", was very low. However, after 2014 this number has been increasing.

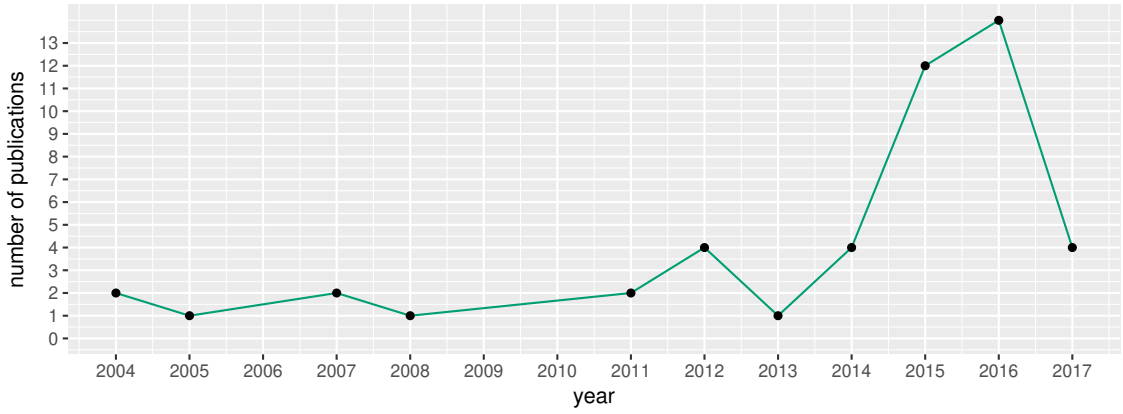

Figure 2.2: Evolution of the number of publications per year from the "Computer Science and Engineering" set (N = 44).

Regarding the decrement of articles in 2017, we should bear in mind that the collection of new documents stopped in May, 2017.

2.3.1.5 Documents publication

From the total of 49 documents in the set of "Computer Science and Engineering" we found 37 different venues. The publication platforms with more than one occurrence in our collection are presented in the Table 2.6.

Table 2.6: Most used platforms for publication of documents from "Computer Science and Engineering".

Platform for publication	n
ArXiv	6
International Conference on World Wide Web	2
Master Thesis	2

ArXiv is an open-access repository of electronic preprints, and it is the more common platform for publication of hate speech. One factor that can be contributing for this, is that hate speech detection is a recent area with work conducted autonomously.

Besides, the results in this subsection point out that the documents found are not presented in venues specific for hate speech. However we try to find if such platforms exist (e.g. journals or conferences). We discovered some conferences more related with hate speech automatic detection (table 2.7), that seem to be in an early stage.

Table 2.7: Conferences related to hate speech detection, respective area of conference and reference.

Conferences related to hate speech detection	Area	Ref
ALW1: 1st Workshop on Abusive Language Online	Computer science	[ACL17]
Workshop on Online Harassment	Computer science	[CHI17]
Text Analytics for Cybersecurity and Online Safety	Computer science	[20117]
Hate speech Conference	Social Sciences	[Hat17]
UNAOC #SpreadNoHate	Social Sciences	[oCU17]
Interdisciplinary conference on Hate Speech	Humanities	[CON17]

2.3.1.6 Number of citations

We computed the number of citations for each document in Google Scholar and concluded that the majority of the works is cited less than four times (Figure 2.3). The top five papers with more citation in our sample are presented in Table 2.8.

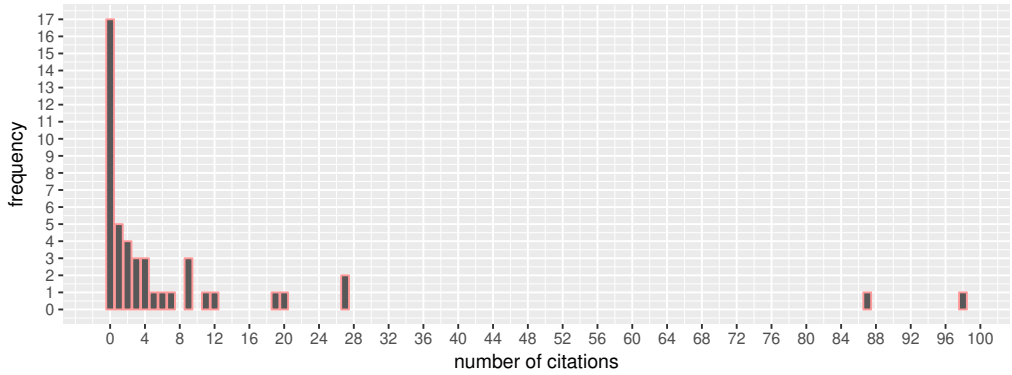

Figure 2.3: Number of citations of the papers from “Computer Science and Engineering”.

2.3.1.7 Keywords in the document

All the keywords referred in the documents from the area “Computer Science and Engineering” were grouped and analysed for absolute frequencies (Table 2.9). We can infer that these documents study hate speech when it is related with:

- **“related concepts”** (cyberbullying, cyber hate, sectarianism and freedom of speech).

Table 2.8: Most cited papers from the "Computer Science and Engineering" set.

Paper	**Citations**	**Reference**
Modelling the Detection of Textual Cyberbullying	87	[DRL11]
Perverts and sodomites: Homophobia as hate speech in Africa	63	[Red02]
Classifying racist texts using a support vector machine	27	[GS04]
Improved Cyberbullying Detection Using Gender Information	27	[DdJOT12]
Detecting Hate Speech on the World Wide Web	20	[WH12a]

- **"machine learning"** (classification, sentiment analysis, filtering systems and machine learning).

- **"social media"** (internet, social media, social network, social networking and hashtag).

Table 2.9: Keywords of the papers from "Computer Science and Engineering".

Keyword	Frequency
cyberbullying	5
social media	5
classification	4
internet	4
freedom of speech	3
hate speech	3
machine learning	3
nlp	3
sentiment analysis	3
social network	3
social networking (online)	3
cyber hate	2
filtering systems	2
hashtag	2
sectarianism	2

2.3.1.8 Social networks

Several articles found usually analyse datasets with messages that were collected from social networks. Twitter is the most commonly used source, followed by general sites, Youtube and Yahoo! (Table 2.10).

2.3.1.9 Number of used instances

Regarding the number of instances per dataset, this number has a wide range of magnitudes (Figure 2.4). Nevertheless, we can conclude that the majority of papers use between 1.000 and 10.000 instances.

Table 2.10: Social networks used in the papers from “Computer Science and Engineering”.

Social network	Frequency
Twitter	16
sites	5
Youtube	3
Yahoo! finance	2
American Jewish Congress (AJC) sites	1
Ask.fm	1
blogs	1
documents	1
Facebook	1
formspring.me	1
myspace.com	1
Tumblr	1
Whisper	1
white supremacist forums	1
Yahoo news	1
Yahoo!	1

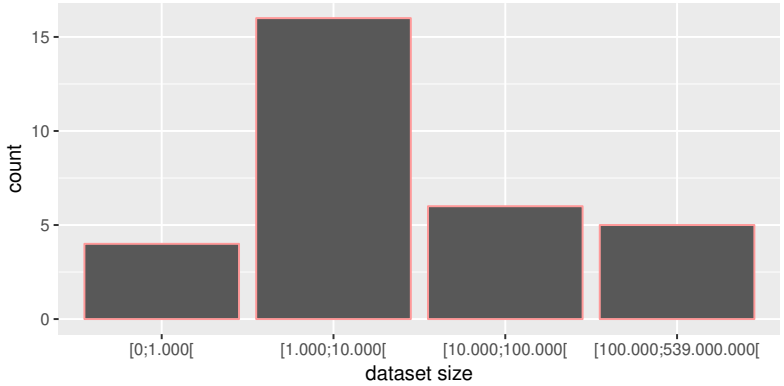

Figure 2.4: Dataset sizes used in the papers from “Computer Science and Engineering”.

2.3.1.10 General or particular hate speech

This subsection aims to analyse if the documents we found focus in general hate speech or in more particular types of hate. The majority (N = 26) considers general hate speech (Table 2.11), however, there is a large number of papers (N = 18) that focus particularly on racism.

2.3.1.11 Algorithms used

The most common approach found is our Systematic Literature Review consists in building a Machine Learning model for hate speech classification. We also found that the most common algorithms used are SVM, Random Forests and Decision Trees (Table 2.12).

Table 2.11: Type of hate speech analysed in the papers from "Computer Science and Engineering".

Hate type	Frequency
general hate speech	26
racism	18
sexism	6
religion	4
anti-semitism	1
nationality	1
other	1
physical/mental handicap	1
politics	1
sectarianism	1
socio-economical status	1

Table 2.12: Algorithms used in the papers from "Computer Science and Engineering".

Algorithms	Frequencies
SVM	10
Random forests	5
Decision trees	4
Logistic regression	4
Naive bayes	3
Deep learning	1
DNN	1
ensemble	1
GBDT	1
LSTM	1
non supervisioned	1
one-class classifiers	1
skip-bigram model	1

2.3.1.12 Type of approach in the document

In order to understand how hate speech is being studied in the "Computer Science and Engineering" articles, we classified the approach of the documents in one the following categories: "algorithms for hate speech", "algorithms but not for hate speech", "descriptive statistics about hate speech", "descriptives statistics but not about hate speech" and "theoretical". In Figure 2.5 we can see that the most common types are "algorithms for hate speech" and "algorithms but not for hate speech". On the other hand, the category "descriptives statistics but not about hate speech" only has one paper about hashtags usage as a way of monitoring discourse [GL15].

In the next subsection we focus on the "algorithms for hate speech" (N = 17) and "descriptives statistics about hate speech" (N = 9) papers, because we want to develop our research in this particular field of hate speech automatic detection.

2.3.2 Documents focusing on descriptives statistics about hate speech detection

In the previous sections we already saw that according to the definitions of hate speech its targets are groups or individuals based on theirs specific attributes, such as ethnic origin, religion, disabil-

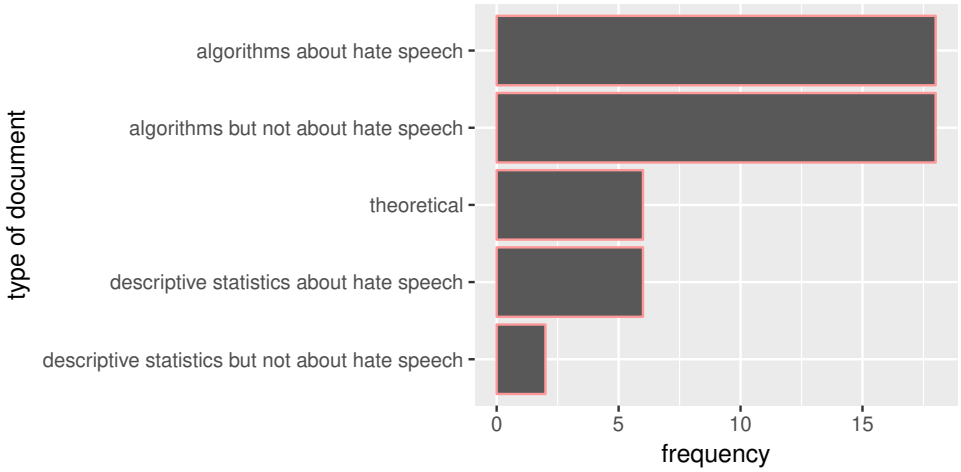

Figure 2.5: Type of papers from "Computer Science and Engineering".

ity, gender identity, age, veteran status, sexual orientation or other. Studies have been conducted in order to describe hate speech online and which groups are more threatened. This section presents the main conclusions found in these articles, that we label as having a more descriptive approach in the problem of hate speech detection. We found descriptive articles about Racism, Sexism, Prejudice towards refugees, Homophobia and General hate speech.

Racism In one study [KW13], the authors tried to understand when hate speech occurs and why messages in social networks are catalogued as racism. They concluded that in the majority of the cases (86%) this is because of the "presence of offensive words". Other motives are "references to painful historical contexts" and "presence of stereotypes or threatening". In order to describe racism across the United States, in another study [Zoo12] the authors try to understand the geographic distribution of racist tweets. They used the information gathered in Twitter to describe the frequencies of tweets in the several states using the geographic reference of the messages [Zoo12].

Sexism In a very simplistic approach, tweets using offensive words towards woman were collected using the Twitter search API. Approximately 5,500 tweets were gathered and coded by one researcher, using a simple binary model. Despite the limitations of the study (e.g. many of the tweets were repeating the title or lyrics from popular songs, with the searched offensive words), it was still relevant for understanding that offensive communication towards woman is a reality in Twitter [HTB16]. A second study also describes misogynistic language on Twitter [BNP+14]. The main conclusions from this study is that 100,000 instances of the word rape used in UK-based Twitter accounts were found, from which around 12% appeared to be threatening. Moreover, approximately 29% of the rape tweets appeared to use the term in a casual or metaphorical way. On the other hand, this study also points out that women are as almost as likely as men to use offensive

terms against women on Twitter. They also found out that the offensive terms are used in a casual or metaphorical way.

Prejudice towards refugees Other study was focused on the annotation of a dataset in German, for hate speech against refugees [RRC+17]. The main goal of this study was to point out the difficulties and challenges when annotating a dataset.

Homophobia Some other study, with an ethnographic methodology, was conducted in Africa [Red02]. Data was collected from several sources (e.g. newspapers, sites). The study concluded that homophobic discourses were using arguments related with Abnormality, Xenophobia, Racism, Barbarism, Immorality, Unpatriotism, Heterosexism, AntiChristianity, UnAfrican, Animalistic behaviour, Inhumane, Criminality, Pathology and Satanism.

General hate speech Finally, other studies take into consideration several types of hate speech at same time. In one particular case [SMC+16] social networks (Twitter and Whisper) were crawled with expressions that follow a rigid pattern:

- I < intensity >< userintent >< hatetarget >

One message following this pattern would be "I really hate people". After collecting the messages, the researchers tried to infer the target of hate in the tweets. With this method they concluded that "race", "behavior" and "physical" were the most hated categories. Finally, in another study an analysis of data recorded by the FBI in 2015 [FBI15] for victims in the USA of single-bias hate crime incidents, showed that the offender's bias was towards different targets in different proportions (Figure 2.6).

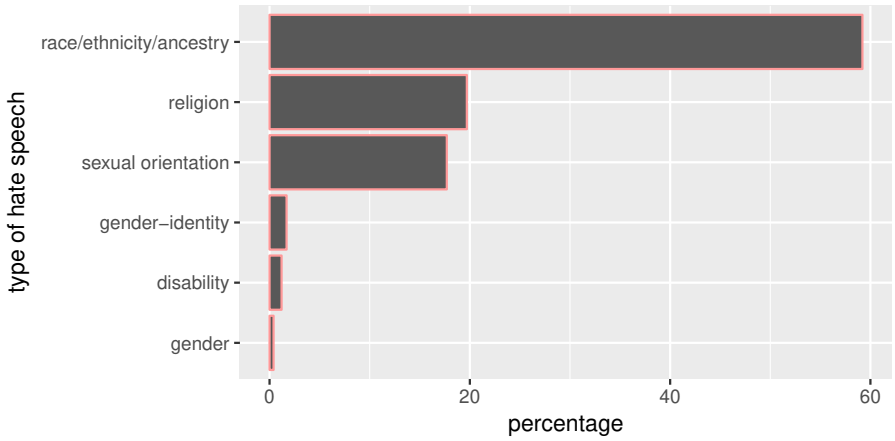

Figure 2.6: Percentage for each type over all hate crimes in USA.

2.3.3 Documents focusing on algorithms for hate speech detection

From our collection of documents, the papers focusing on "algorithms for hate speech detection" are the most important in our study. This is the case because we aim to research in this specific topic. First, in what concerns to the methodology followed in these papers, the researchers used machine learning for hate speech classification. Additionally, in the majority of the works the used language is English. However there were some exceptions. In these cases the considered languages were Dutch [THL+16], German [RRC+17], or Italian [DVCD+17]. In the next sections we present with more detail how these studies obtain datasets, the main authors from the papers and we make a comparison in the performances of the different approaches.

2.3.3.1 Datasets used in the papers

In the majority of the 17 papers focusing in "algorithms for hate speech", new different data was collected and annotated. However in only a few studies data is made available for other researchers ("own, available"), and in only one case an already published dataset is used ("published dataset") (Figure 2.7). In these circumstances it is more difficult to compare the approaches in the different works.

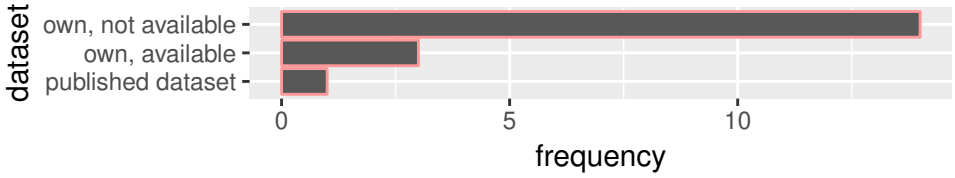

Figure 2.7: Dataset availability in the documents with "algorithms about hate speech".

The datasets available in the area are described in the subsection 2.4. We also focus in the procedure for data collection and annotation in subsection 3.3.

2.3.3.2 Main authors

We could also conclude that we have a high number of distinct authors from these papers (Table 2.13).

2.3.3.3 Achieved performances

In the collected papers, several metrics were computed in order to estimate the performance of the models. Precision, Recall and F-measure were the most common metrics and in some other studies Accuracy and AUC (Area Under Curve) were also considered. In Table 2.14, the results of the studies are presented in descending order of the F-measure value. These results should be analysed with some caution, because in the several papers different configurations, datasets, and definitions were compared and we try to summarize here the best results for each paper.

Table 2.13: List of the author in the papers with "algorithms about hate speech".

Author	Frequency
Pete Burnap	3
Edel Greevy	2
Matthew L. Williams	3
Alan f. Smeaton	1
Ben Verhoeven	1
Chikashi Nobata	1
Dirk Hovy	1
Elise Lodewyckx	1
Heng Xu	1
Irene Kwok	1
Jing Zhou	1
Joel Tetreault	1
Julia Hirschberg	1
Lisa Hilte	1

We concluded that it is not clear to understand which approaches perform better. On one hand, the best results were achieved when deep learning was used. However, this was not a consistent result, because in some other studies the performance of deep learning was much worse. In the following sections we focus in the usage of text mining for hate speech or related concepts detection.

2.3.4 Text mining approaches in automatic hate speech detection

The study of automatic hate speech detection has been growing only in the few last years. However some studies have already been conducted as we presented previously. Papers focusing in algorithms for hate speech detection, and also other studies focusing on related concepts (e.g. Cyberbullying), can give us insight about which features to use in this classification task.

Finding the right features to tackle a problem can be one of the more demanding tasks when using machine learning. Therefore we allocate this specific section to describe the features already used by other authors. We divide the features into two categories: general features used in text mining, that are common in other text mining fields; and the specific hate speech detection features, that we find only in the hate speech detection task and are intrinsically related with the characteristics of this problem. We present our analysis in this section.

2.3.4.1 General features used in text mining

The majority of the papers we found try to adapt strategies already known in text mining to the specific problem of automatic hate speech detection. We define general features as the features commonly used in text mining. We start by the most simplistic approaches, that use dictionaries and lexicons.

Dictionaries One strategy in text mining is the use of dictionaries. This approach consists in making a list of words (the dictionary) that are searched and counted in the text. These frequencies

Table 2.14: Results evaluation of the papers in the metrics Accuracy (Acc), Precision (P), Recall (R), F-measure (F) and AUC, respective features and algorithms used.

Year	Acc	P	R	F	AUC	Features	Algorithms	Paper
2017	-	0.93	0.93	0.93	-	-	Logistic Regression, Random Forest, SVM, GBDT, DNN, CNN	[BGGV17]
2004	-	∼0.90	0.9	0.9	0.9	BOW, n-grams, POS	SVM	[Gre04]
2017	-	0.91	0.9	0.9	-	TF-IDF, POS, sentiment, hashtags, mentions, retweets, URLs, number of characters, words, and syllables	logistic regression, SVM	[DWMW17]
2017	-	0.833	0.872	0.851	-	POS, sentiment analysis, word2vec, CBOW, n-grams, text features	SVM, LSTM	[DVCD+17]
2016	-	0.83	0.83	0.83	-	n-grams, lenght, punctuation, POS	skip-bigram model	[NTT+16]
2014	-	0.89	0.69	0.77	-	n-gram, typed dependencies	Random Forest Decision Tree, SVM	[BW14]
2015	-	0.89	0.69	0.77	-	n-gram, typed dependencies	Random Forest Decision Tree, SVM, Bayesian Logistic Regression, ensemble	[BW15]
2016	-	0.72	0.77	0.73	-	user features	logistic regression	[WH16]
2016	-	0.79	0.59	0.68	-	BOW, dictionary, typed dependencies	SVM, Random forest Decision Tree	[BW16]
2015	-	0.65	0.64	0.65	-	rule-based approach, sentiment analysis, typed dependencies	non supervised	[GZDL15]
2012	-	0.68	0.6	0.63	-	template based strategies, word sense disambiguation	SVM	[WH12a]
2016	-	0.49	0.43	0.46	0.63	dictionaries	SVM	[THL+16]
2015	-	-	-	-	0.8	paragraph2vec	logistic regression	[DZM+15]
2016	0.91	-	-	-	-	word2vec	deep learning	[YWX16]
2013	0.76	-	-	-	-	n-grams	Naive bayes	[KW13]
2016	-	0.73	0.86	-	-	Topic Modelling, Sentiment Analysis, Tone Analysis, Semantic Analysis, Contextual Metadata	one-class classifiers, Random Forest, Naive bayes, Decision Trees	[AS17]
2004	-	0.93	0.87	-	-	BOW, n-grams, POS	SVM	[GS04]
2014	-	0.97	0.82	-	-	TF-IDF, n-grams, topic similarity, sentiment analysis	Naive bayes	[LF14]

can be used directly as features or to compute scores. In the case of hate speech detection this has been conducted using:

- Content words (such as insult and swear words, reaction words, personal pronouns) collected from www.noswearing.com [LF15].
- Number of profane words in the text, with a dictionary that consists of 414 words including acronyms and abbreviations, where the majority is adjectives and nouns [DdJOT12].
- Label Specific Features which consisted in using frequently used forms of verbal abuse as well as widely used stereotypical utterances [DRL11].
- Ortony Lexicon was also used for negative affect detection; the Ortony lexicon contains a list of words denoting a negative connotation and can be useful because not every rude comment necessarily contains profanity and can be equally harmful [DRL11].

This methodology can be used with an additional step of normalization, by the total number of words in the comment [DdJOT12]. Besides, it is also possible to use this kind of approach with regular expressions [Mal14].

Distance Metric Some studies have pointed out that in the offensive text messages it is possible that the offensive words are obscured with an intentional misspelling, often a single character substitution [WH12a]. Examples of this terms are "@ss", "sh1t" [NTT+16], “nagger”, or homophones, such as “joo” [WH12a]. The Levenshtein distance can be used with this purpose, and it is defined as the minimum number of edits necessary to transform one string into the other [NS15]. The distance metric can be used to complement dictionary-based approaches.

Bag-of-words (BOW) Another model similar to dictionaries is the use of bag-of-words [BW16, KW13] [GS04]. In this case, a corpus is created based on the words that are in the training data, instead of a pre-defined set of words, as in the dictionaries. After collecting all the words, the frequency of each one is used as a feature for training a classifier. The disadvantages of this kind of approaches is that the word sequence is ignored, and also its syntactic and semantic content. Therefore, it can lead to misclassification if the words are used in different contexts. To overcome this limitation n-grams were implemented.

N-grams n-grams are one of the most used techniques in hate speech automatic detection and related tasks [BW16, NTT+16, WH16, LF14, BYH+, GS04, DWMW17, BGGV17]. The most common n-grams approach consists in combining sequential words into lists with size N. In this case, the goal is to enumerate all the expressions of size N and count the occurrences of them. This allows to improve classifiers’ performance because it incorporates at some degree the context of each word. Instead of using words it is also possible to use n-grams with characters or syllables. This approach is not so susceptible to spelling variations as when words are used. In a study character n-gram features proved to be more predictive than token n-gram features, for the specific problem of abusive language detection [MT16].

However, n-grams also have disadvantages. One disadvantage is that related words can have a high distance in a sentence [BW16] and a solution for this problem, such as increasing the N value, slows down the processing speed [Che11]. Also, studies point out that higher N values (5) perform better than lower values (unigrams and trigrams) [LF14]. In a survey [SW17], researchers report that n-grams features are often reported to be highly predictive in the problem of hate speech automatic detection, but perform better when combined with others.

Profanity Windows Profanity windows are a mixture of a dictionary approach with n-gram. The goal is to check if a second person pronoun is followed by a profane word within the size of a window and then create a boolean feature with this information [DdJOT12].

TF-IDF The TF-IDF (term frequency-inverse document frequency) was also used in this kind of classification problems [DRL11]. TF-IDF is a measure of the importance of a word in a document within a corpus and increases in proportion to the number of times that a word appears in the document. However, it is distinct from a bag of words, or n-gram, because the frequency of the term is off-setted by the frequency of the word in the corpus, which compensates the fact that some words appear more frequently in general (e.g. stop words).

Part-of-speech Part-of-speech (POS) approaches make it possible to improve the importance of the context and detect the role of the word in the context of a sentence. These approaches consist in detecting the category of the word, for instance, personal pronoun (PRP), Verb non-3rd person singular present form (VBP), Adjectives (JJ), Determiners (DT), Verb base forms (VB). Part-of-speech has also been used in hate speech detection problem [GS04]. With these features it was possible to identify frequent bigram pairs, namely PRP_VBP, JJ_DT and VB_PRP, which would map as "you are" [DRL11]. It was also used to detect sentences such as "send them home", "get them out" or "should be hung" [BW14]. However, POS proved to cause confusion in the classes identification [BW14], when used as features.

Lexical Syntactic Feature-based (LSF) In a study [Che11], the natural language process parser, proposed by Stanford Natural Language Processing Group [Gro17], was used to capture the grammatical dependencies within a sentence. The features obtained are pair of words in the form "(governor, dependent)", where the dependent is an appositional of the governor (e.g. "You, by any means, an idiot." means that "idiot", the dependent, is a modifier of the pronoun "you," the governor). These features are also being used in hate speech detection [Che11].

Rule based approaches Some rule-based approaches have been used in the context of text mining. A class association rule-based approach, more than frequencies, is enriched by linguistic knowledge. Rule-based methods do not involve learning and typically rely on a pre-compiled list or dictionary of subjectivity clues [HL14]. For instance, rule-based approaches were used to classify antagonistic and tense content on Twitter and they used associational terms as features. They

also included accusational and attributional terms targeted at only one or several persons following a socially disruptive event as features, in an effort to capture the context of the terms used.

Participant-vocabulary consistency (PVC) In a study about cyberbullying [RH16], this method is used to characterize the tendency of each user to harass or to be harassed, and the tendency of a key phrase to be indicative of harassment. For applying this method it is necessary a set of messages for the same user. In this problem for each user, it is assigned a bully score (b) and a victim score (v). For each feature (e.g. n-grams) a feature-indicator score (w) is used. It represents how much the feature is an indicator of a bullying interaction. Learning is then an optimization problem over parameters b, v, and w.

Template Based Strategy The basic idea of this strategy is to build a corpus of words, and for each word in the corpus, collect K words that occurring around [Pow11]. This information can be used as context.

This strategy has been used for feature extraction in the problem of hate speech detection as well [WH12a]. In this case a corpus of words and a template for each word was listed, as in:

- template literal "W-1:go W+0:back W+1:to"

This is an example of a template for a two word window on the word "back".

Word Sense Disambiguation Techniques This problem consists in identifying the sense of a word in the context of a sentence, when it can have multiple meanings [Yar94]. In a study, the stereotyped sense of the words was took into consideration, in order to understand if the text is anti-semitic or not [WH12a].

Typed Dependencies Typed dependencies were also used in hate speech related studies. First, to understand the type of features that we can obtain with this, the Stanford typed dependencies representation provides description of the grammatical relationships in a sentence, that can be used by people without linguistic expertise [DMM08]. It was used for extracting Theme-based Grammatical Patterns [GZDL15] and also for detecting hate speech specific othering language [BW15, BW14], that we will present within the specific hate speech detection features. Some studies report significant performance improvements in hate speech automatic detection based on this feature [BW16, GZDL15].

Topic Classification With these features the aim is to discover the abstract topic that occurs in a document. In a particular study [AS17], topic modelling linguistic features were used to identify posts belonging to a defined topic (Race or Religion).

Sentiment Bearing in mind that hate speech has a negative polarity, authors have been computing the sentiment as a feature for hate speech detection [LF14, LF15, GZDL15, DWMW17, AS17, DVCD+17]. Different approaches have been considered (e.g. multi-step, single-step) [SW17]. Authors usually use this feature in combination with others which proved to improve results [LF14].

Deep Learning Deep learning techniques are recently being used in text classification and sentiment analysis, with high accuracy [YWX16]. Some authors [DZM+15] use a paragraph2vec approach to classify language on user comments as abusive or clean and also to predict the central word in the message [DZM+15]. FastText is also being used [BGGV17]. A problem that is referred in hate speech detection is that in that case sentences must be classified and not words [SW17]. Averaging the vectors of all words in a sentence can be a solution, however this method has limited effectiveness [NTT+16]. Alternatively, other authors propose comment embeddings to solve this problem [DZM+15].

Other features Other features also used in this classification task were based in techniques such as **Named Entity Recognition** (NER) [CH15], **Topic Extraction** [LF14], Word Sense Disambiguation Techniques to check **Polarity** [NTT+16, GZDL15], frequencies of **personal pronouns** in the first and second person, the presence of **emoticons** [DdJOT12, DVCD+17] and **capital letters** [DdJOT12]. Before the feature extraction process, some studies have also used **stemming** and removed **stop-words** [LF14, BW14, DWMW17].

Characteristics of the message were also considered such as hashtags, mentions, retweets, URLs, number of tags, terms used in the tags, number of notes (reblog and like count) and link to multimedia content such as image, video or audio attached to the post [AS17].

2.3.4.2 Specific hate speech detection features

Complementary to the approaches commonly used in text mining analysis, several specific features are being used to tackle the problem of hate speech automatic detection. We briefly present the features found.

Othering Language Othering has been used as a construct surrounding hate speech [BW16] and consists in analysing the contrast between different groups by looking at "Us versus Them". It describes "Our" characteristics as superior to "Theirs" which are inferior, undeserving and incompatible [DAKAA15]. Expressions like "send them home" show this kind of cognitive process. Othering terms and language were identified using an implementation of the Stanford Lexical Parser, along with a context-free lexical parsing model, to extract typed dependencies [BW16]. Typed dependencies provide a representation of syntactic grammatical relationships in a sentence. For instance [BW14], in the tweet "Totally fed up with the way this country has turned into a heaven for terrorists. Send them all back Home", one resultant typed dependency is nsubj(home--5, them-2). This identifies the relationship nsubj, which is an abbreviation of nominal subject

between the fifth word 'home' and the second word 'them'. The association between both words, is an example of "othering" phrase, because the opposition between "them" from "us", through the relational action of removing "them" to their "home" [BW14].

Perpetrator Characteristics Some other studies also consider features more related with the social network graph. In this particular case [WH16], this study was linking the available messages from a same user and focusing on the user characteristics like gender and geographic localization.

Objectivity-Subjectivity of the language On one hand, some authors [GZDL15] argue that hate speech is related with more subjective communication. In this study they use a rule-based approach to separate objective sentences from subjective ones and, after this step, they erase the objective sentences from their analysis. However, other authors [WH12a] point out that in some other cases prejudiced and hateful communication can be conducted recurring to scientifically worded essays. In this case, in some sites they found that the anti-semitic speech was not presenting explicitly pejorative terms. Instead, it presented extremely anti-semitic ideologies and conclusions in a scientific manner. The differences found in both studies cited in this subsection point out that hate speech detection can occur in several forms. Therefore it is important to understand what is contributing for its different expressions and how to include the plurality of the concept and several nuances in the developed model.

Declarations of superiority of the ingroup In addition to the question of the objectivity and subjectivity of the language, declarations of superiority of the ingroup can also be considered hate speech. In this case, hate speech can also be conducted when there are only defensive statements or declarations of pride, rather than attacks directed toward a specific group [WH12a].

Focus on particular stereotypes In some studies [WH12a] authors hypothesize that hate speech often employs well known stereotypes and therefore they subdivide such speech according to the stereotypes. This approach can be useful because each stereotype has a specific language: words, phrases, metaphors and concepts. For instance, anti-Hispanic speech might make reference to border crossing; anti-African American speech often references unemployment or single parent upbringing; and anti-Semitic language often refers to money, banking and media [WH12b]. Given this, creating a language model for each stereotype is a necessary prerequisite for building a general model for all hate speech [WH12b]. In some other studies [SMC+16] authors also point out the different hate speech categories. They combine Hatebase along with the categories reported by FBI for hate crimes and they ended up with nine categories: Race, Behavior, Physical, Sexual orientation, Class, Gender, Ethnicity, Disability, Religion and "other".

Intersectionism of oppression Intersectionality is a concept that points out the connection between several particular types of hate speech (e.g. when burka is prohibited it can be analysed either as an islamophobic or sexist behavior, because this symbol is used by muslims, but just for

woman). Intersectionism of several kinds of oppressions presents a particular challenge for the automated identification of hate speech and it has been considered in literature. In a study [BW16] the intersectionism is considered only in the evaluation of the model, where more than one class was regarded at the same time and not in the feature extraction process.

2.3.4.3 Summary from the text mining approaches

In conclusion, in this subsection we tried to understand which specific features have been used in hate speech detection. The different studies use different features, and in some cases the conclusions seem contradictory in the comparison of the different studies. The results of the categorization conducted are summarized in the two Figures (2.8 and 2.9).

2.3.5 Main conclusions from the Systematic Literature Review

We conducted a Systematic Literature Review with the goal to understand the state of the art and opportunities in the field of automatic hate speech detection. This proved to be not an easy task, mostly because this topic has been widely discussed in other fields, such as social sciences and law, and therefore we find a large number of documents that would require more resources to process. In order to solve this problem, we focused only in the documents from computer science and engineering, and we concluded that the number of articles focusing on hate speech has been increasing in the last years. At the same time that this field is growing, it is possible to notice that it remains in an early phase. The existing papers are published in a wide range of venues, not specific for hate speech, and the few conferences towards this topic that exist are having now its first editions. Besides, the majority of the papers found also has a low number of citations.

Regarding the practical work conducted, hate speech is being analysed in connection with other related concepts (e.g. Cyberbullying), social media and machine learning. From all the possible approaches from machine learning, hate speech automatic identification is being tackled as a classification task. The wide majority of the studies considers this a binary classification problem (hate speech vs. not hate speech). However, a few have also used a multiclass approach, where racism is one of the classes more regarded.

In the majority of the works, researchers collect new datasets. Twitter is the preferred social network, and English the most common language. We concluded that the authors do not use public datasets, and do not publish the new ones they collect. This makes very difficult to compare results and conclusions. Comparative studies and surveys are also scarce in the area. Finally, regarding the features used, we observed that the majority of the studies consider general approaches of text mining and do not use particular features for hate speech.

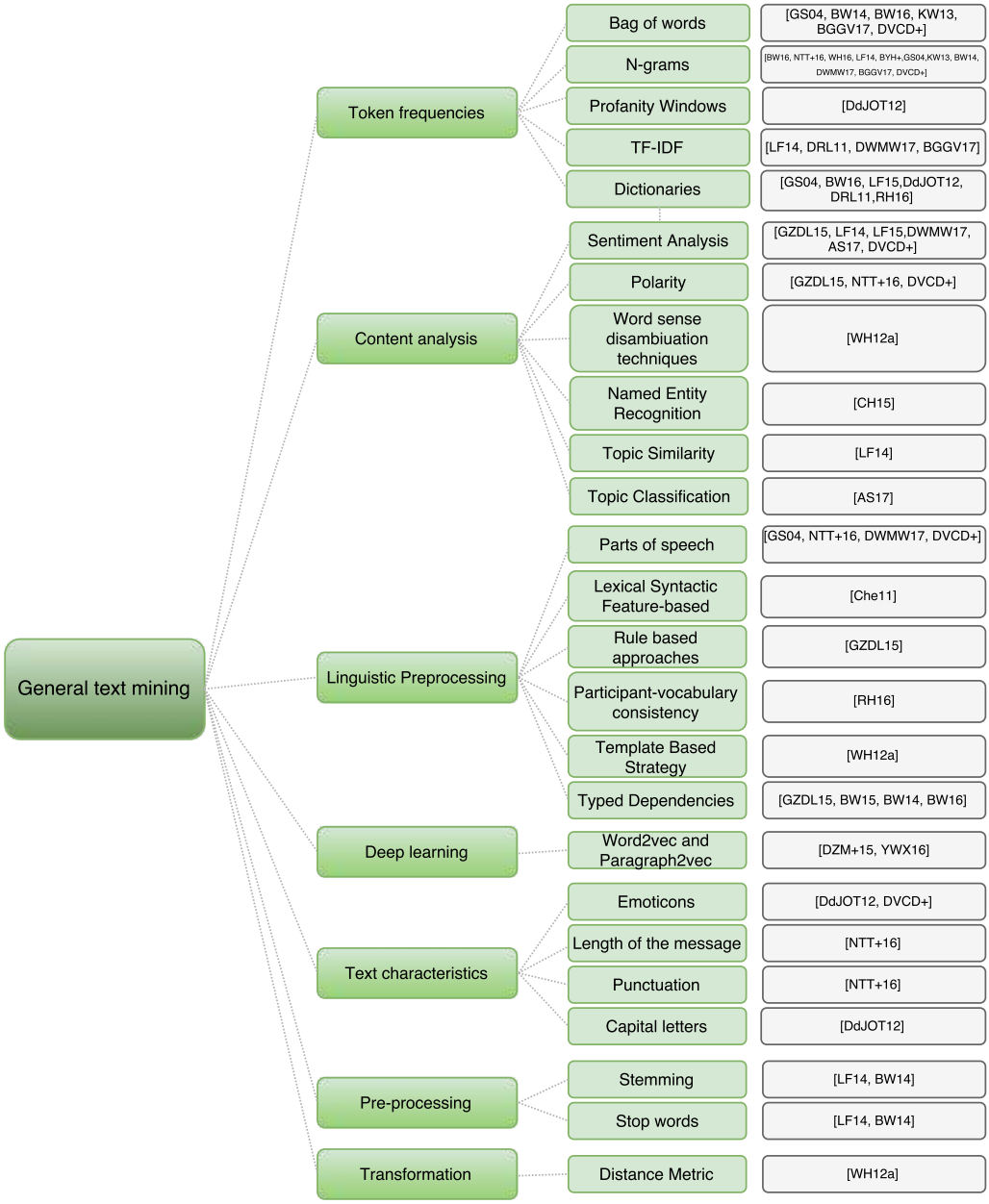

Figure 2.8: Papers using generic text mining features.

.

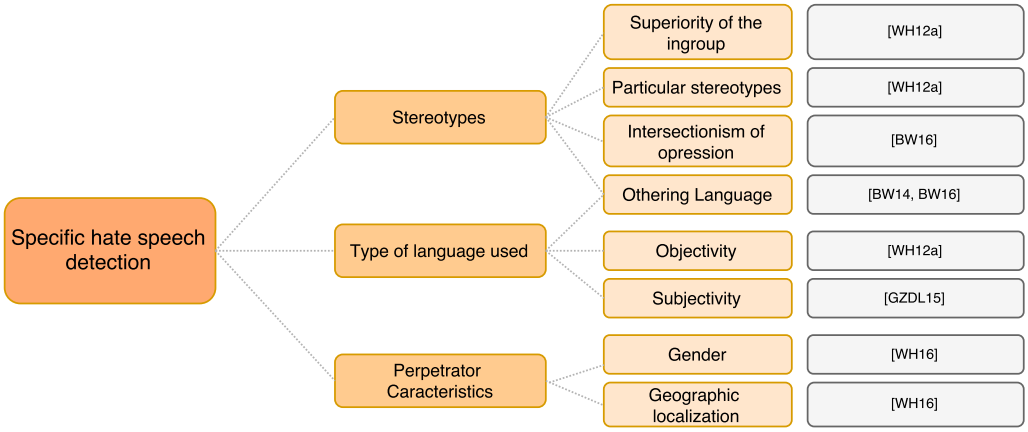

Figure 2.9: Papers using specific hate speech detection features .

2.4 Data useful for hate speech classification

In the conducted literature review some datasets and corpus were found. We present these datasets in Table 2.15.

Despite the fact that some datasets and corpus for hate speech already exist there is no official established one yet.

2.5 Open source projects for hate speech automatic detection

In order to check if there are any projects available for hate speech automatic detection that can be used as examples or sources for annotated data, we also inspected GitHub using the keyword "hate speech" in the available search engine. We found 25 GitHub repositories with some content. We describe here the main conclusions we achieved from this search.

2.5.1 The type of approach

We manually classify the type of approach followed in the projects (Figure 2.10) and we concluded that the majority of the projects is concerned about classifying messages as hate speech. Some other projects are also concerned with collecting messages with hate speech (crawling) and building dictionaries to help in the task of hate speech detection. Besides, some projects were also relating hate speech with sentiment evaluation of messages and latent semantic analysis.

In what concerns to the programming languages, all projects were developed in Python, except three developed with JavaScript and Java.

2.5.2 Datasets used in the GitHub projects

The datasets used in the GitHub projects are analysed regarding its source, language and availability. In what concerns to the source, the majority of projects works with Twitter, and all the

Table 2.15: Datasets and corpus for hate speech detection.

Source	Name	Distribution	Year	Type	Number of instances	Classes Used	Language	Link to the article
University of Copenhagen	Hate Speech Twitter annotations	GitHub repository	2016	Dataset	16914	sexist, racist	English	[Was16]
CrowdFlower	Hate speech identification	available for the community	2015	Dataset	14510	offensive with hate speech, offensive with no hate speech, not offensive	English	[Cro17]
Yahoo Webscope	Abusive language dataset	Not available	2016	Dataset	2000	hate speech, not offensive	English	[Yah17]
User-Centred Social Media Graduiertenkolleg	German Hatespeech Refugees	Creative Commons Attribution-ShareAlike 3.0 Unported License	2016	Dataset	470	hate speech, not offensive	German	[UCS16]
Hatebase	Hatebase	available for the community	2017	Corpus	-	-	Universal	[Kag13]
Hades	Hades	available for the community	2016	Corpus	-	-	Dutch	[CLi16]
-	Hate speech and offensive language	available for the community	2017	Corpus	-	-	English	[Dav17]

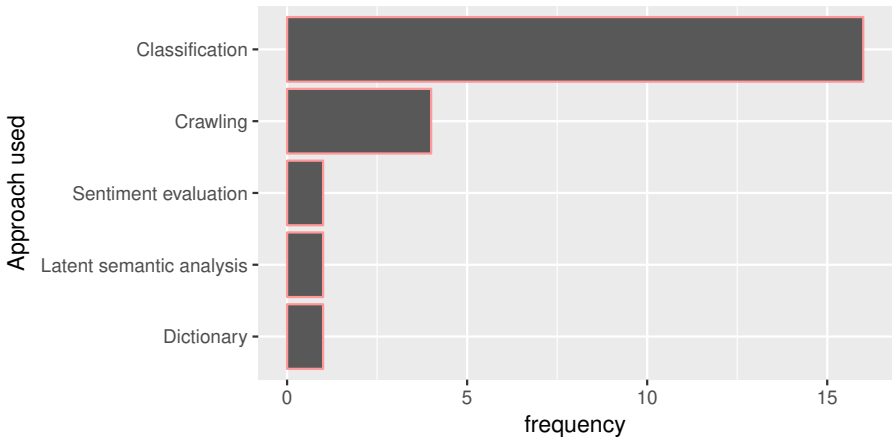

Figure 2.10: Approaches used in open source projects about hate speech.

projects use messages in English, except two that use Dutch [THL+16] or Finnish. These findings are congruent with our Systematic Literature Review. Regarding availability, we were interested

in access to new datasets or messages already annotated. We saw that two repositories provide some more datasets [Eys16, Tho16b]. However we concluded that the majority of the projects does not provide any new data and there is also a few projects that use datasets already described in the Section 2.4.

2.6 Difficulties in detecting hate speech

In this overview of the topic of hate speech automatic detection we already discussed that this is not an easy task. Other researchers also point out some difficulties inherent to this task:

- Low agreement (33%) in hate speech classification by humans, indicating that this classification would be harder for machines [KW13].
- The task of annotating a dataset is also more difficult because it requires expertise about culture and social structure [RH16].
- The evolution of social phenomena and language makes it difficult to track all racial and minority insults [NTT+16]. Besides, language evolves quickly mainly among young populations [RH16] that communicate frequently in social networks.
- Despite the offensive nature of hate speech, abusive language may be very fluent and grammatically correct, can cross sentence boundaries and it is also common the use of sarcasm in it [NTT+16].
- Finally, hate speech detection is more than simple keyword spotting [NTT+16].

We find relevant to present these difficulties here in order to bear in mind the kind of challenges we can have in our research.

2.7 Opportunities in the field of automatic hate speech detection

It is also important to point out that the systematic literature review conducted allowed us to spot some opportunities in this field. We present it in the next paragraphs.

- **Open source platforms or algorithms** - in our Systematic Literature Review we found that the documents describe methods, features extracted and algorithms used. However it is rare to find works with open source code, and also no open source tools are available for hate speech automatic detection.
- **Definition of a main dataset** - the definition of a main dataset would be important in this field in order to make easier the comparison between the different studies in the field.
- **Comparative studies** - Studies comparing the different models in the field are also missing.

- **Research mainly in English** - As we described previously, the majority of the studies focus in English. Besides, only isolated studies were conducted for German, Dutch and Italian. In this case, research in other languages commonly used on the internet is also needed (e.g. French, Mandarin, Portuguese, Spanish).

We can conclude that any research that either makes available open source platforms or algorithms, helps to define a main dataset, comparative studies and research in other languages besides English will be helpful within this field.

2.8 Conclusions from the overview on hate speech automatic detection

In this overview of automatic hate speech detection research we tried to clarify concepts and understand how this field has been evolving over the last years. For the concept of hate speech we concluded that it has been defined in several platforms, from social networks to other organizations, and therefore different definitions for the same concept exist. However, in order to build a model for hate speech automatic detection, we need a clear and unique definition and we tried to do it in this section. We also concluded that it is easier to understand hate speech when we compare it with other related concepts, such as hate, cyberbullying, abusive language, discrimination, profanity, toxicity and flaming. More than this, the research in these other concepts can bring us insight about how to tackle the problem of hate speech detection in text.

Additionally, in order to have a picture from the state of the art in the field, we conducted a Systematic Literature Review. We concluded that the number of studies and papers published in automatic hate speech detection in text is limited and usually those works regard the problem as a machine learning classification task. In this field, researchers tend to start by collecting and classifying new messages, and often the used datasets remain private. This slows down the progress in this research field because less data is available and also makes more difficult to compare the results in the different studies. Nevertheless, we found three available datasets, in English and German.

Regarding the features used in these studies, we found that both general text mining approaches, and specific hate speech features are used. For the first, those are mainly n-grams, POS, rule based approaches, sentiment analysis and deep learning features, such as word2vec. For the specific hate speech detection features, we found mainly othering language, superiority of the ingroup, focus on stereotypes. Regarding the journals, conferences or workshops in the area, we found they are rare and in an initial phase. This points out that this is an area still in the beginning.

Finally, in this chapter we also spotted some opportunities in the field, such as the lack of open source platforms that automatically classify hate speech; no comparative studies that would summarize the approaches conducted so far; and, because the majority of the research was conducted only in English, languages such as French, Mandarin, Portuguese or Spanish have no advances in this area.

Based on the conclusions from the overview on the topic, we decided that our thesis should focus in research for Portuguese. In the next chapter we present how we collect data for this language and we describe the dataset obtained.

Chapter 3

Hate Speech Dataset Annotation for Portuguese

After we conducted an overview on the topic of hate speech automatic detection in text, we reached some conclusions that guide our approach in this thesis. First, in the previous chapter we concluded that there is lack of research in automatic hate speech detection for Portuguese and also a lack of annotated datasets in this language. We decided then that one goal of this thesis should be to fill this gap, because annotated data is essential in machine learning for supervised tasks. Other conclusion that we achieved is that the majority of the studies we found collects new data but does not make it available for other researchers. This is a limitation because it makes more difficult to compare approaches and replicate findings. In these cases it is desirable to use the same data.

One final concern is related with the concept of hate speech. In the previous section we presented hate speech as a complex concept and we want to incorporate in our approach the nuances of this phenomena. For that, we take into consideration that different subtypes of hate speech exist and that at the same time these intersect each other. We summarize our goals for this chapter as:

- Annotate a dataset for Portuguese (with messages from Brazil and Portugal).
- Find an annotation method that considers the existence of several subtypes of hate speech and preserves the intersectional nature of these classes.

There are different ways of obtaining trained datasets and in our work we use one of the most common methods. In this case messages are collected in social networks and manually annotated. However, it is very important that the annotation procedure assures quality in the dataset, because it can have an impact in the model performance. In the following sections we present how we tackled this problem. First we present hate speech classification as a problem with hierarchical classes. We then analysed the methodologies presented in other dataset annotation studies, as a base for our annotation. We describe our annotation procedure and the results from some statistics of the achieved dataset in terms of n-grams and POS.

3.1 Hate speech classification as a problem with hierarchical classes

With our study we aim to find an annotation method that considers the particularities of hate speech. First, instead of a singular phenomena, some previous studies [WH12b] defend the existence of several subtypes of hate speech (e.g. racism, sexism, homophobia). This approach can be useful because each subcategory of hate speech has specific words, phrases, metaphors and concepts. For instance, anti-Hispanic speech is more related with border crossing; while anti-African American often references unemployment or single parent upbringing; and anti-Semitic language refers more to money and banking. Given this, creating a language model for each stereotype is a necessary prerequisite for building a general model for all hate speech [WH12b].

At the same time, intersectionality is a concept that points out to the connection between the particular subtypes of hate speech. Intersectionism of several kinds of oppressions presents a particular challenge for the automated identification of hate speech [BW16]. In this dataset annotation we want to preserve both ideas: that different subtypes of hate speech exist and that these classes intersect each other. We propose a hierarchical classification approach for our representation.

3.2 Hierarchical classification

Hierarchical classification is opposed to flat classification, where the predefined categories are treated in isolation and there is no structure defining the relationships among them [DC00]. On the other hand, in the case of hierarchical classification there is a structure defining the hierarchy between the classes. In this case one or more classifiers are constructed at each level of the category tree and each classifier works as a flat classifier at that level. Each instance will be classified until it reaches a final category, that can be a leaf or an internal category [HCT07]. The authors distinguish two basic cases: a tree structure, where each class has one parent classes, except the root; and a directed acyclic graph structure, where more than one parent can occur. We present both in the following paragraphs.

(Virtual) category tree In the virtual category tree structure, each category can belong to at most one parent. Two different configurations can occur: the virtual category tree and the category tree structure. In the virtual category tree documents can only be assigned to the leaf categories [DC00]. Other configuration is the category tree structure, where documents can also be assigned into both internal and leaf categories [WZH01]. One example, is the Binary Hierarchical Classifier where the tree has a number of leaf nodes equal to the number of classes in the output space (Figure 3.1).

(Virtual) directed acyclic category graph In the virtual directed acyclic category graph structure, categories are organized as a Directed Acyclic Graph (DAG) where a class can have more than one parent. Similar to the case of category tree, the difference between virtual and non-virtual

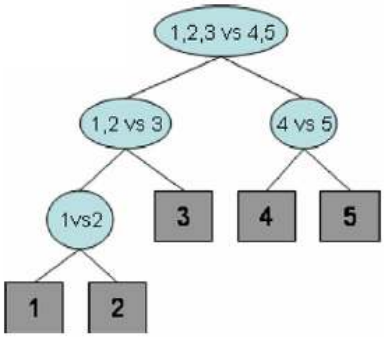

Figure 3.1: An example of Binary Hierarchical Classifier architecture which consists of N - 1 classifiers arranged as a binary tree (Image from [HCT07]).

directed acyclic category graph is that documents can only be assigned to leaf categories in the first and also to internal nodes in the second [HCT07] (Figure 3.2).

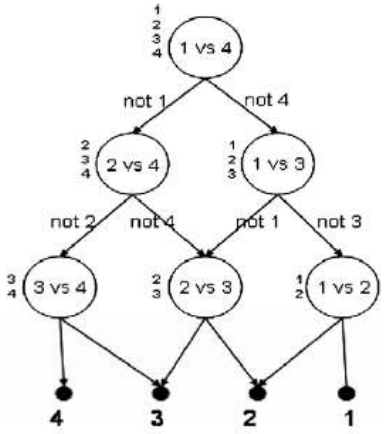

Figure 3.2: An example of DAGSVM architecture which uses a rooted binary directed acyclic graph with $n(n - 1)/2$ internal nodes and n leaves (Image from [HCT07]) .

3.2.1 Hate speech classes representation

As we already presented in Chapter 2, hate speech is a complex phenomena and its identification is not an easy task. In order to better adapt the process of classification to the nature of hate speech and also to regard the intersectional dimension of it [BW16], we adopted an hierarchical class representation. We propose that a directed acyclic category graph (nonvirtual) is the structure that best fits this problem [HCT07]. For representing hate speech we propose a graph of classes with the following properties:

- "Hate speech" class corresponds to the root of the graph, because all the messages from the positive class should be marked as hate speech.
- The lower nodes of the graph inherit the classes from the upper nodes up to the root.
- A lower node can have more than one parent.
- If hate speech can be divided in several types of hate, several nodes descend from the root node. This originates a second level (Table 3.1) of classes according to the targets of the hate (e.g. Racism, Homophobia, Sexism).
- This second level of nodes can also be divided into subgroups of targets. For instance, racist messages can be targeted against black people, chinese people, latin people or other.
- The division of the classes can continue until we find no more distinct groups.

Table 3.1: Subtypes of hate speech definition.

Class	**Definition**
Sexism	Hate speech based on gender. Includes hate speech against, for instance woman.
Body	Hate speech based on body, such as fat, thin, tall or short people.
Origin	Hate speech based on the place of origin. Includes hate speech against Mexican, for instance.
Homophobia	Hate speech based on sexual orientation.
Racism	Hate speech based on ethnicity.
Ideology	Hate speech based on the people's way of thinking, such as feminism, left wing ideology.
Religion	Hate speech based on the religion.
Health	Hate speech based on the health condition, such as against disabled people.
Other-Lifestyle	Hate speech based on life habits, such as vegetarianism.

We present here part of the graph that we propose (Figure 3.3) and an example of an instance from the leaf node (Table 3.2). The complete graph used is also presented in the Appendix A.

Table 3.2: Hate speech example.

Message	**Class**
Thanks, fat ugly woman.	Fat, ugly, women

In this case, the message from Table 3.2 would inherit some of the labels from the sample chart, namely: "Hate speech", "Sexism", "Hate based on the body", "Hate against women", "Hate against fat people" and "Hate against woman". Considering hate speech classification as a problem with a directed acyclic category graph structure allows us to transform a variable class as a set of binomial dummy classes, where each message can belong or not to all the possible classes in the graph. This approach has several advantages.

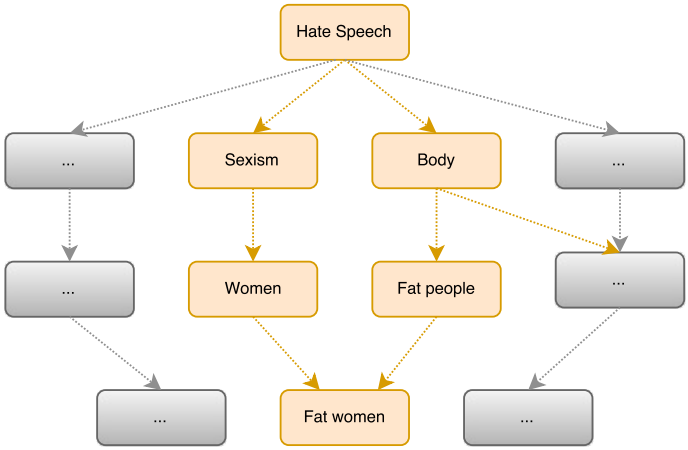

Figure 3.3: Hate speech classes represented with a directed acyclic category graph structure.

The main advantage of this classification system is that it describes better the relationship between the different types of hate speech, that are not isolated from each other. With this approach we can also preserve very rare classes, that can be new forms of hate speech for instance. At the same time, we can take rare hate speech subtypes as part of a bigger class (e.g. use a message to build a model for predicting sexism even if the message was catalogued as hate against fat women). Besides, with this approach we are providing a dataset where we aim to enumerate all the possible classes for each message. This makes possible to study each subtype of hate speech individually, or in relation to the others, depending on the goals of future studies. In the next section we focus on articles where processes of dataset annotation for hate speech were conducted.

3.3 Methodologies presented in other studies

Some datasets for hate speech classification were already collected in previous studies. In this section we aim to present the methodologies followed, in order to have a background for planning and conducting a study for Portuguese. We analyse these studies regarding three different aspects: the procedure for messages collection, the annotation method and the main conclusions of the study.

3.3.1 Hate speech in Twitter dataset

The first article providing a Hate Speech dataset and the respective annotation procedure is from 2016 [WH16]. We present here how they collected and annotated messages and the main conclusions achieved in this study.

Messages collection In this study an initial manual search was conducted in Twitter in order to collect common slurs and terms pertaining to religious, sexual, gender, and ethnic minorities. In

the results, the authors identified frequently occurring terms in tweets that contain hate speech and used these terms to look for messages.

Annotation method The main researcher of the article, together with a gender studies student, manually annotated the dataset. They followed a set of guidelines and considered a tweet offensive if: it uses a sexist or racial slur; attacks a minority; seeks to silence a minority; criticizes a minority (without a well founded argument); promotes, but does not directly uses, hate speech or violent crime; criticizes a minority and uses a straw man argument; blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims; shows support of problematic hashtags; negatively stereotypes a minority; defends xenophobia or sexism; contains a screen name that is offensive; the tweet is ambiguous (at best); and the tweet is on a topic that satisfies any of the above criteria. They annotated a total of 16,914 tweets: 3,383 of that for sexist content sent by 613 users, 1,972 for racist content sent by 9 users, and 11,559 for neither sexist or racist and sent by 614 users.

Main conclusions The inter-annotator agreement had a Cohen's Kappa of 0.84 when they considered two classes (racism and sexism). They made the dataset available as pairs "tweet identification" and "class label" at GitHub. The authors of the study found that n-gram provide good results in this task and adding demographic information has little improvement in the performance of the classification model. From the demographic features, gender was the one with better results.

Dataset properties In this study two persons annotated a dataset with 16,914 tweets for English, using the classes "Racism", "Sexism" and "Neither".

3.3.2 Yahoo Webscope abusive language dataset

Another article from 2016 [NTT+16] describes a dataset where messages are classified in the general class "abusive language", and within the subclasses "hate speech", "derogatory" and "profanity". Despite that they end up not providing the annotated dataset, the conducted procedure is still of interest in the context of our work.

Messages collection The authors sampled 2,000 comments posted on Yahoo! Finance and News.

Annotation method A Mechanical Turk experiment was conducted for annotating the messages. In this study the workers from Mechanical Turk were provided with some annotation instructions and informations. First hate speech was defined as language which attacks or demeans a group based on race, ethnic origin, religion, disability, gender, age, disability, or sexual orientation/gender identity. They also provided some examples to the workers: "The JEWZ are Market Manipulators Hate them freaking Jews !! To bad Hitler did not finish them off !!!"; "I hate the mall because there are too many unemployed apefricans and mudsharks."; "300 missiles are cool! Love

to see um launched into Tel Aviv! Kill all the gays there!"; "EGYPTIANS ARE DESCENDED FROM APES AND PIGS. NUKE EGYPT NOW Ah me tu se ku sh". The 2000 comments were annotated.

Main conclusions Considering binary classification (only with the class "abusive language"), the workers had a Fleiss's Kappa of 0.401. However when using the fine-grained three classes (hate speech, derogatory and profanity) the Fleiss's Kappa dropped to 0.213. The authors experimented several new syntactic features and embedding features, and found them to be successful when combined with standard NLP features. n-grams performed good in this dataset, as well.

Dataset properties In this study they annotated a dataset with 2000 messages in English, using the classes "Hate Speech", "Derogatory", "Profanity" and "Neither".

3.3.3 Hate speech against refugees in German dataset

The majority of the studies that we found for hate speech were conducted for English. However, some other languages were considered. The study we present here is an example of a dataset collection and annotation in German, in the specific topic of hate speech against refugees and it was also published in 2016 [RRC+17].

Messages collection In the procedure of collecting messages, Twitter was used as a source of data. They used 10 hashtags that can be used for insult and offence. With these hashtags the authors gathered initially 13,766 messages, however they excluded retweets or replies, as they can be hard to understand without the rest of the conversation. In addition, they removed duplicates and near-duplicates. Search terms related with hate speech but not with refugees were also discarded and also the tweets containing pictures or links. The authors ended up with 470 tweets.

Annotation method The 541 tweets were split into six parts and each part was annotated by two out of six annotators in order to determine if hate speech was present or not. The annotators were rotated so that each pair of annotators only evaluated one part. The offensiveness of a tweet was rated also on a 6-point Likert scale.

Main conclusions In this study they found that the inter-annotator agreement was low, with a Krippendorff's alpha of 0.38, even among researchers familiar with the definitions. The results of this study pointed out that hate speech is a vague concept that requires definitions and guidelines in order for having reliable annotations. They also provided solutions for improving this classification task. First they referred that collecting multiple labels for each tweet can be an advantage. Moreover, the authors also note that considering hate speech detection as a regression problem, instead of a binary classification task, can also improve the classifier's performance.

Dataset properties In this study they annotated a dataset with 470 messages in German, using only the class "Hate Speech".

3.3.4 CrowdFlower Hate Speech identification

The most recent paper in hate speech dataset annotation dates from 2017. This paper presents data collected using the CrowdFlower platform [DWMW17]. We describe this study in the next paragraphs.

Messages collection The researchers in this study began with a hate speech lexicon compiled by Hatebase.org in English. Then, using the Twitter API they searched for tweets containing the terms from the initially computed lexicon. From these tweets, the authors reached a total of 33,458 Twitter users, from whom they extracted the timeline for each user, resulting in a set of 85.4 million tweets. Finally, from this corpus they took a random sample of 25,000 tweets containing terms from the lexicon and had them manually coded by CrowdFlower workers.

Annotation method Three or more annotators in CrowdFlower viewed short text segments and identified if it: contains hate speech; is offensive but without hate speech; or is not offensive at all. The annotators were provided with a definition along with a paragraph explaining hate speech in further detail. Users were asked to think not just about the words appearing in a given tweet but about the context in which they were used. They were instructed that the presence of a particular word, despite being offensive, did not necessarily indicate a tweet is hate speech. They have used the majority decision in CrowdFlower for each tweet to assign a label. Some tweets were not assigned labels as there was no majority class. This resulted in a sample of 24,802 labelled tweets.

Main conclusions The intercoder-agreement score provided by CrowdFlower was 92% and a total percentage of only 5% of tweets were coded as hate speech by the majority of coders. Consistent with previous work, these study pointed out that certain terms are particularly useful for distinguishing between hate speech and offensive language. Besides, the results also illustrate how hate speech can be used in different ways: it can be directly send to a person or group of people targeted; it can be espoused to nobody in particular; and it can be used in conversation between people.

Dataset properties In this study they annotated a dataset with 14,510 messages in English, using only the class "Hate", "Offensive" or "Neither"

3.3.5 Summary of the annotated datasets

As a conclusion for this section we summarize and compare the methodologies followed in the four studies presented previously (Table 3.3).

Table 3.3: Summary of the annotated datasets presented in literature.

Dataset **Criteria**	**Hate Speech Twitter annotation [WH16]**	**Yahoo Webscope Abusive Language Dataset [NTT+16]**	**German Hate speech Refugees [RRC+17]**	**CrowdFlower Hate Speech identification [DWMW17]**
Messages collection	Twitter search engine with frequently occurring terms in hate speech tweets.	Sampled from comments posted on Yahoo! Finance and News	Twitter search engine with 10 hashtags that can be used in an insulting or offensive way.	Twitter search engine with words from Hatebase.org.
Number of messages	16,914 tweets	2,000 comments from Yahoo! Finance	541 tweets	24,802 tweets
Annotation method	Manually total of 2 annotators.	Mechanical Turk experiment.	The 541 tweets were split into six parts and each part was annotated by two out of six annotators.	Three or more annotators in CrowdFlower.
Do they have annotation guidelines	Yes	Yes	Not described. Used a Likert scale.	Yes
Annotators agreement	Cohen's Kappa of 0.84	Binary class Fleiss's Kappa of 0.401.	Krippendorff's alpha of 0.38	intercoder-agreement score provided by CrowdFlower of 92%.
Dataset availability	Yes. Tweet IDs and labels at Github.	No.	Yes. Tweet ID's and text available	Yes. Tweet ID's and text available in CrowdFlower dataset
Language	English	English	German	English
Main conclusions	- using n-gram provided a solid foundation for the features. - demographic information gives little improvement to the model.	- The authors experimented several new features combined with standard NLP features that worked. - n-grams performed good in this dataset.	- Hate speech is a vague concept that requires definitions and guidelines. - Collecting multiple labels for each tweet can work. - Consider hate speech detection as a regression problem, instead of a binary classification task, can also work.	- Some terms are particularly useful for distinguishing between hate speech and offensive language. - Hate speech can be used directly with the targets; it can be espoused to nobody in particular; and it can be used in conversation between people.

In our study we consider the procedures to collect other datasets presented in this section as a guideline. We then adapted this procedure to the context of Portuguese language. We describe our dataset annotation methodology in the next section.

3.4 Dataset Annotation in Portuguese

In our Systematic Literature Review presented in Chapter 2, we concluded that with the exception of Dutch, German and Italian, there is no significant research being done in other languages than English. Therefore there is lack of research and annotated data in Portuguese for hate speech detection as well. We find important to tackle this problem also in Portuguese, because this language is in the top 5 more used in Twitter [Fox13]. We present here the procedure that we followed for message collection, message annotation, the final dataset transformation and analysis.

3.4.1 Phase 1 - Messages collection

In the procedure of message collection we bear in mind some principles. First, in this dataset annotation we aim to have a higher proportion of hate speech messages comparing to previous research. Other studies have found that it is difficult and costly to assure a corpus with equal proportion of hateful and harmless comments. This is the case because hateful messages are more rare than benign comments and therefore a large number of messages have to be annotated to find a considerable number of hate speech instances [SW17]. The referred proportions are 5% [DWMW17] and 11.6% [BW15] of hate messages in the respective datasets. In our study, we aim to increase these percentages, in order to minimize the annotation effort and costs.

One second aspect is that we have a preference for finding spontaneous occurrences of hate speech and avoid bias in our sampling process. In some previous studies we have noticed that researchers look for messages using expressions already known as linked to hate speech [WH16, DWMW17]. However, if this increases the proportion of hate speech messages, at the same time it has the disadvantage of focusing the resulting data in specific topics and certain subtypes of hate speech [SW17]. This practice perpetuates that the model will learn towards the knowledge that the researcher already has, and we also think that it makes more likely that new existent forms of hate speech are omitted.

Finally we aim to have messages from a diverse source of users, and both in Portuguese from Brazil and from Portugal. In the next subsections we present in detail the steps followed for the message collection (Figure 3.4).

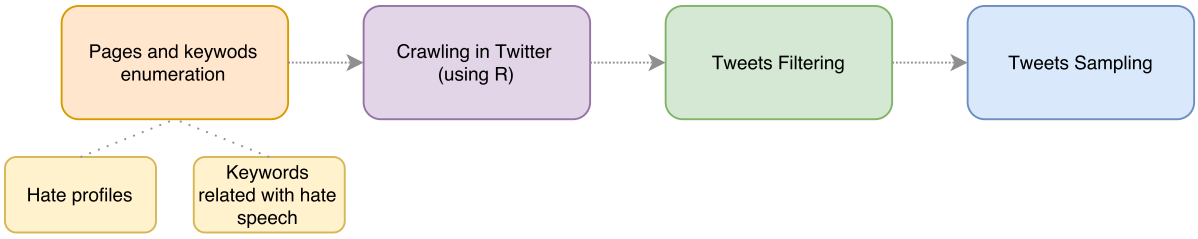

Figure 3.4: Method for messages collection.

3.4.1.1 Pages and keyword enumeration

For the source of our messages we decided to use Twitter, because this social network is used worldwide and provides an accessible and easy to use API. On one hand, it is possible to use the search with words or expressions for collecting messages. However, we can also collect messages from specific profiles as well. In the case of our problem we considered these two possibilities because both methods are complementary. With the first we access to a wider range of tweets from different profiles, but we restrict the discourse with a specific word or expression. On the other hand, collecting messages from profiles allows access to more spontaneous discourses, but from a more restricted number of users.

Hate profiles We started by collecting Twitter profiles known for posting only hate messages in several different topics. We expect that these types of pages have a large amount of hateful messages. One way to reach to these profiles is by using the search engine with words like "hate", "hate speech" or "offensive".

Keywords related with hate speech In order to have more messages, we then enumerated each type of hate that we have found in the literature and tried to find, using the Twitter search engine, keywords, hashtags and profiles where these types of hate speech could be present.

Finally, we used both methods ("Hate profiles" and "Keywords related with hate speech") and we ended up with a list of 29 different profiles, 19 keywords and 10 hashtags that we used in the crawling process (Appendix B).

3.4.1.2 Crawling

After the profiles and keywords enumeration, we made a crawler in R using both the search with keywords and also the collection of all posts from a profile. For the search with the keywords we had to remove accentuation (e.g. #LugarDeMulherÉnaCozinha was not working and it had to be replaced by #LugarDeMulherEnaCozinha). With both gathering methods, using profiles and search words, we then collected a total of 42,930 messages. We conducted the crawling from 8 to 9 of March 2017.

3.4.1.3 Tweets Filtering

We then preprocessed the tweets according to the following rules:

- We only kept tweets that were categorized by Twitter as in Portuguese.
- We checked to assure that no repetitions and no retweets with the same text were kept.
- We remove the HTML.

- We remove tweets that without hashtags, URL and user mentions had less than 3 words. We omit those because we expect that short messages will be marked as hate speech just by the presence of a specific offensive word. That corresponds to a dictionary-based approach and in our study we aim to find more complex examples of hate speech.

After the initial cleaning we got 33,890 tweets available, however we sampled this set due to the limitation in resources for the annotation task.

3.4.1.4 Tweets sampling

In order to reduce our initial sample of 33,890 messages, we search for a criteria to select the messages. We noticed that the search instances returned a number of tweets from different magnitudes (e.g. some profiles had only around 30 messages while others had more than 3,000). We decided then to limit to a maximum of 200 tweets per search instance in order to keep a more diverse source of tweets.

3.4.2 Phase 2 - Messages annotation

Despite the differences between the previous studies that we analysed in Section 2.4, the majority of the described works present instructions for the annotation task. Some authors point out that having vague annotation guidelines [SW17] is a problem for hate speech detection due to the complexity of the task [RRC+17]. In our work we prepared a complete set of annotation instructions.

3.4.2.1 Annotation instructions

In order to better standardize the annotation procedure and to make clear hate speech related concepts, a set of instructions and examples was defined (Appendix C). These are based on the definitions, rules and examples that we presented already in the chapter 2. The annotators were given the instructions, as guidelines in the classification of the messages.

3.4.2.2 Annotation procedure

The dataset was annotated mainly by 1 researcher (thesis author) and additionally a random subset of 500 messages was also annotated by a second annotator not connected to the field of research. The annotators were instructed to use one or more classes from the graph in Appendix A. In order for the annotation task to be easier, the annotators did not have to give all the parent categories (e.g. from the example in the Table 3.2, the annotators would only have to give "Fat women"). Instead an algorithm was used to discover from a children category all the parents until the root ("Fat women" label originates then the labels "Fat people", "Hate speech based on body", "Women", "Sexism" and "Hate speech" as well).

3.4.3 Annotation results

3.4.3.1 Sample description

Our dataset has 5,668 tweets, from which we were able to check the author of 4,548, which corresponds to 1,156 distinct users. The tweets collected have timestamps from 2009-11-25 to 2017-03-09, however, the number of tweets not from January, February and March of 2017 is residual (Figure 3.5).

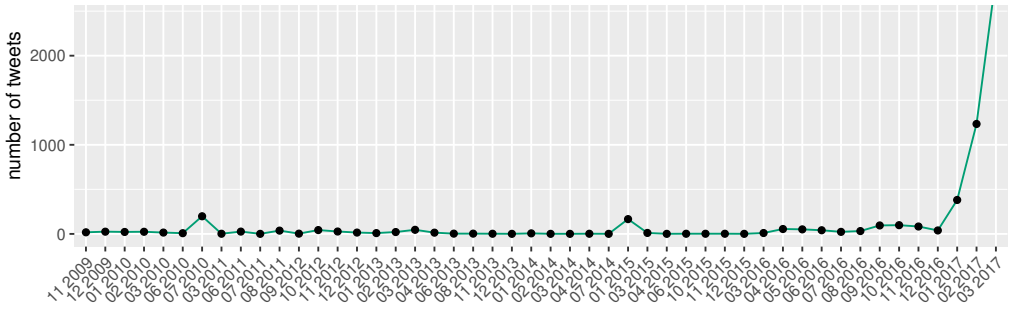

Figure 3.5: Temporal distribution of the collected messages.

3.4.3.2 Agreement between annotators

In order to better understand the agreement between the two annotators in the 500 messages, we used the Cohen's Kappa [GLG+12]. We have tried different configurations.

Agreement in the annotator input without any dataset manipulation In this case, without any manipulation in the classification from the annotators, we are checking if they agree in the classes used and also in the classes order. In other words, we are matching the strings used by both researchers to classify each message. We achieved K = 0.53.

Agreement when the problem is transformed into a binomial classification ("message with hate speech" vs "none") In this case we converted the string input of the annotators to a binomial class ("message with hate speech" vs "none"), which corresponds to a dummy variable "hate speech". This is the root node in the graph of classes (Figure 3.3). In this case we are matching if both annotators agree when hate speech occurs, without considering the type of hate, and we achieved K = 0.58.

Agreement considering hierarchical classes In this case we transformed the annotators string input into a matrix of dummy variables. In this matrix the columns corresponds to each of the types of hate speech in our graph of classes (Appendix A), and each row to a message. When we compare the two matrices, one for each of the annotators, we achieved K = 0.72.

Agreement by class, considering hierarchical classes We can also consider the agreement of the annotators, considering each type of hate speech. We rank the classes by the best agreement and we found that the consensus was higher in hate speech against "lesbians", "based on health", "homophobia" and "disabled people" (Table 3.4). The classes with only one instance in any of the annotators were removed from this analysis.

Table 3.4: Agreement evaluation for each class, with the total positive messages for each class for each annotator.

classes	K	Total annotator 1	Total annotator 2
Lesbians	0.879	59	53
Health	0.856	3	4
Homofobia	0.823	69	61
Disabled people	0.799	2	3
Refugees	0.763	13	13
Migrants	0.751	15	14
Sexism	0.669	134	104
Trans women	0.662	6	9
Men	0.657	12	15
Women	0.642	109	75
Fat women	0.637	30	16
Body	0.637	32	17
Fat people	0.637	32	17
Ideology	0.609	14	15
Feminists	0.581	13	14
Hate speech	0.569	245	213
Racism	0.501	18	13
Religion	0.493	5	11
Black people	0.435	11	7
Origin	0.329	3	3
Islamists	0.329	2	10
Gays	0.300	4	9
Ugly women	0.276	24	4

Regarding the annotators agreement, we concluded that the distinct configurations we tried led to different results and our best solution was when using the hierarchical approach. Despite in this case we have already an acceptable score (K = 0.72), there is still room for improvement. Besides, we also found that for different types of hate speech the agreement can have contrasting values, which points out that some specific types of hate speech can be more difficult to classify than others. We also concluded that in the different studies diverse measures for agreement evaluation were used (Fleiss's Kappa, Krippendorff's alpha, Cohen's Kappa), which makes more difficult to compare this metric in the several studies.

3.4.3.3 Frequencies of types of hate speech

In order to find the most frequent types of hate in the messages collected we plotted the types of hate sorted by frequency (Figure 3.6).

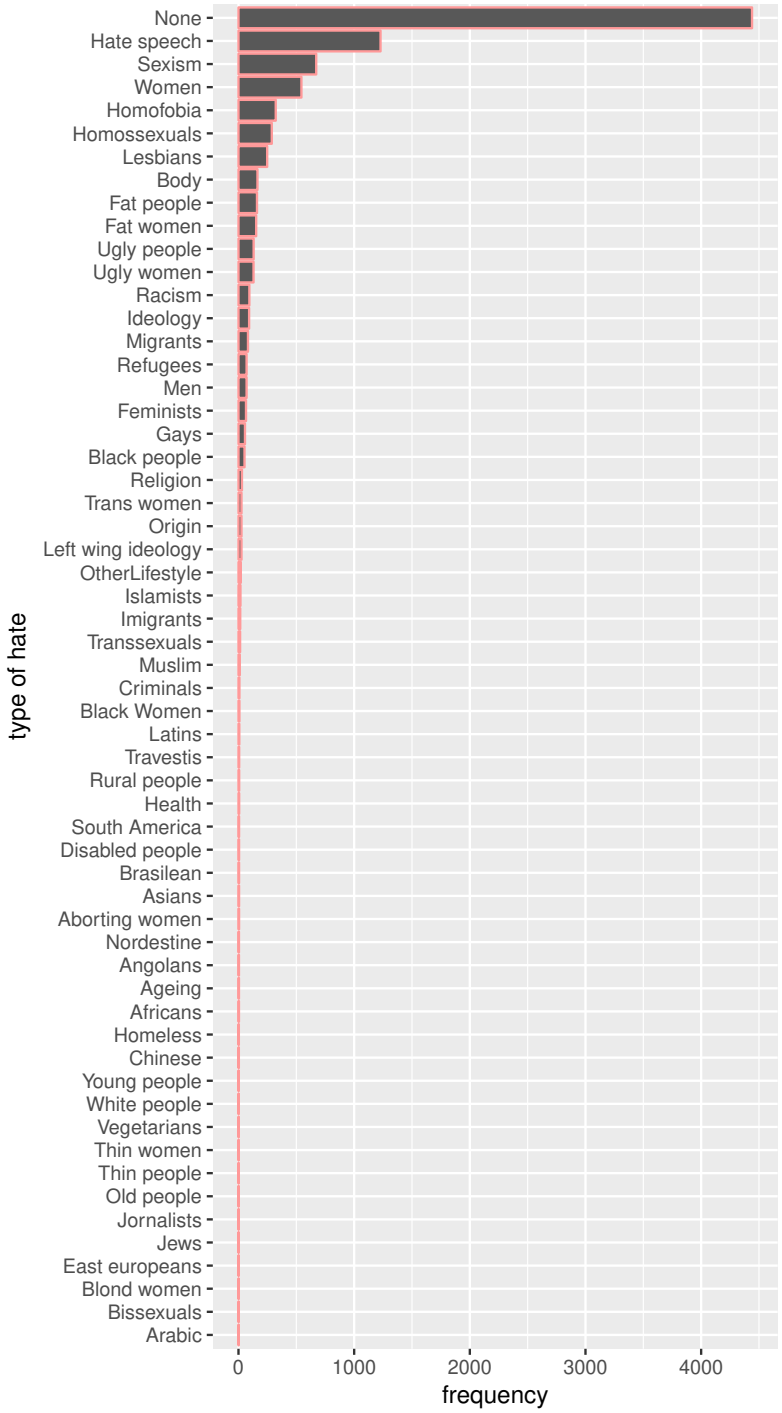

Figure 3.6: Types of hate frequencies in the dataset order by frequency.

We concluded that from the 5,668 messages, 1,228 contains some type of hate speech, which corresponds to 22% of the messages, while 4,440 contains no hate speech. This result proves that we achieved our goal of increasing the percentage of hate messages, comparing to the best case we found in literature, that was 11,6%. Regarding the frequencies of the hate, we can not

infer from our study which types of hate speech are more common in Twitter, because we were using a sample method that is not random, but based on search words and specific profiles instead. For matter of simplicity, and because we have a large number of classes that are uncommon in the dataset we present in the next subsections the results only for the root class ("Hate speech") and the classes that are immediate children of hate speech ("Health", "Homophobia", "Ideology", "Origin", "Racism", "Religion", "Sexism" and "Other-lifestyle").

3.4.3.4 Temporal distribution

We aimed to find if the different types of hate speech are distinct not only in frequency but also in the moments when the messages are published. To investigate this question we took into consideration that the types of hate have different number of messages in our sample (Figure 3.6) and because of that we used relative frequencies. For each hate type we checked the percentage of messages that occurs in each hour of the day (Figure 3.7) and also in each week day (Figure 3.8).

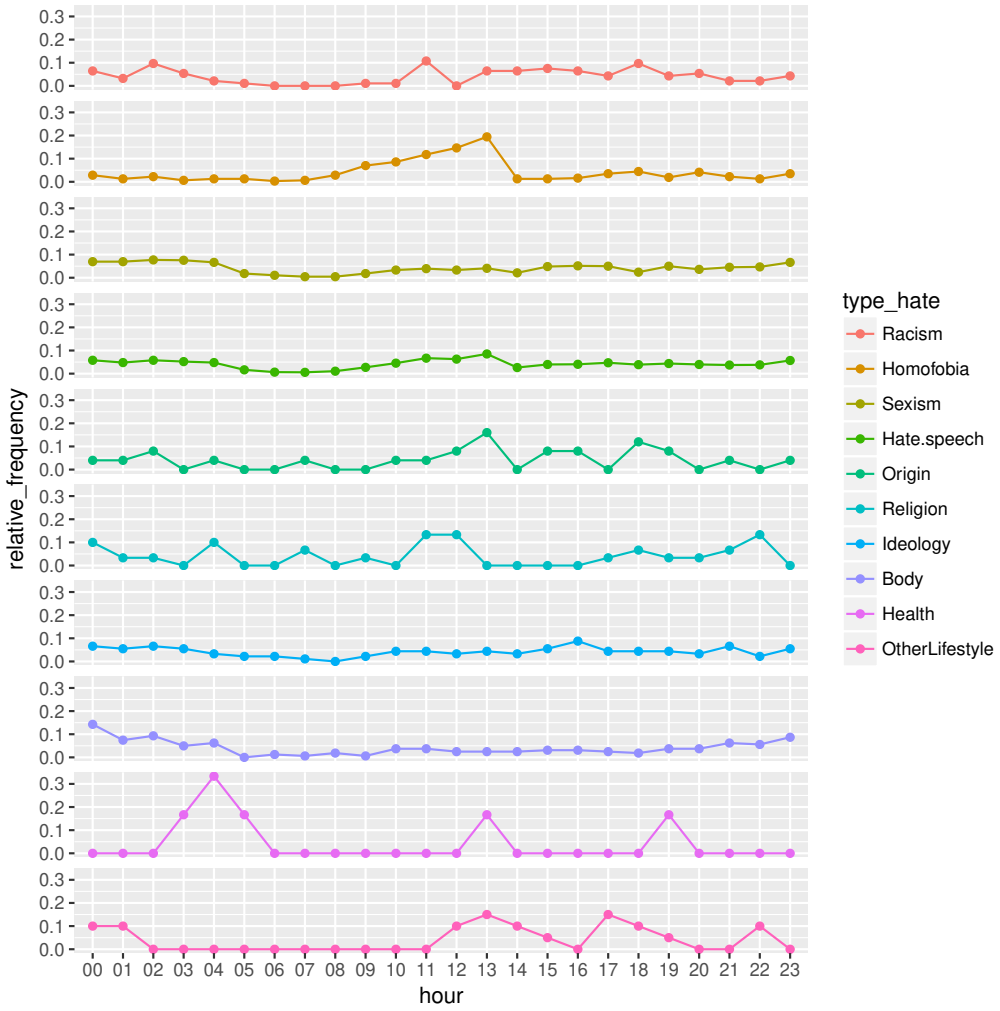

Figure 3.7: Relative frequencies of each type of hate by the day hour.

We found that for the different types of hate speech the hours with more publications are not the same. First “Sexism”, “Religion”, “Origin” and “Ideology” have a homogeneous distribution during the day. However, “Racism” and “Health” are more published during the night (between midnight and 6h), around midday and between 18h and 19h. On the other hand, homophobic (“Homophobia”) messages are more published around midday and, finally, hate messages based on “others lifestyle” are more published between 12h and 24h.

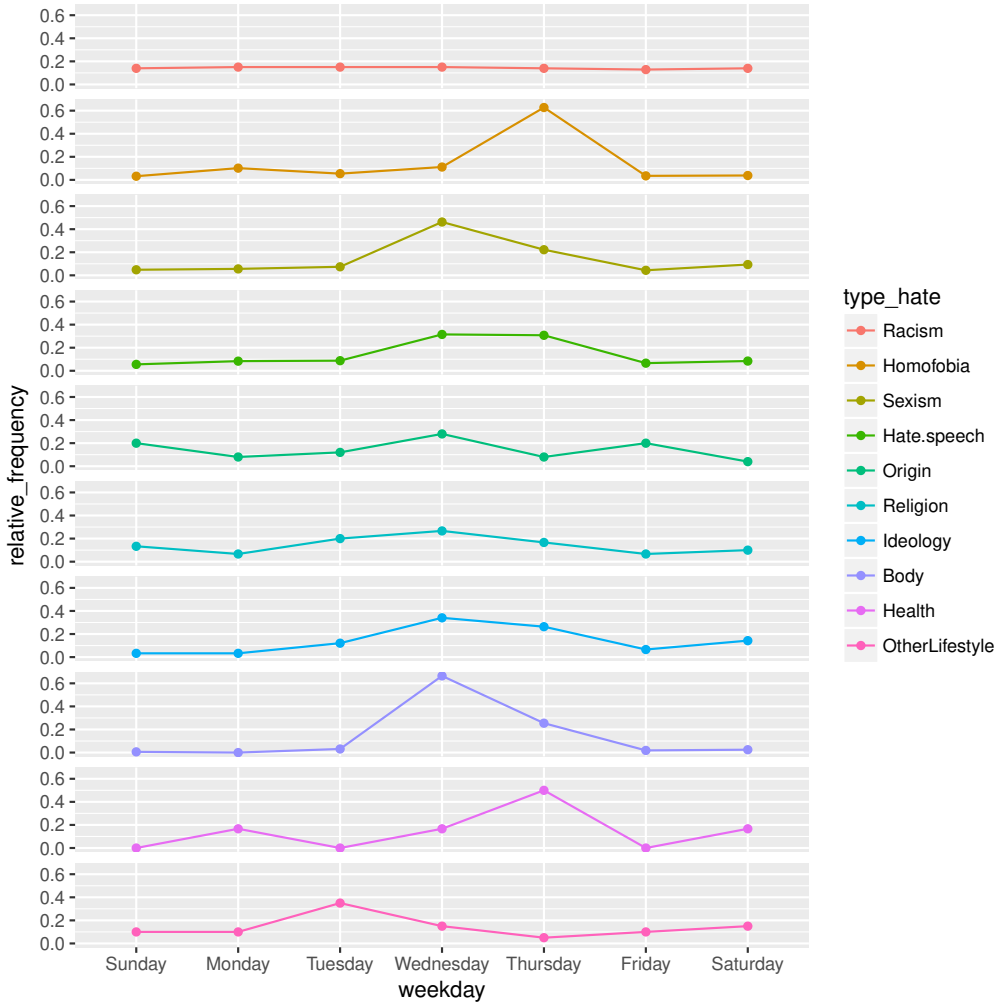

Figure 3.8: Relative frequencies of each type of hate by the week day.

Regarding the distribution of the messages during the week days (Figure 3.8), we can also find different patterns. “Racism”, “Origin” and “Religion” have similar frequencies among all the week days. Besides, the majority of hate speech based on “Others Lifestyle” conditions occurs on Tuesdays; based on “Body” and “Sexism” occurs on Wednesday; and finally based on “Health” and “Homophobia” occurs on Thursday. Finally, Sundays, Mondays, Fridays and Saturdays are days where no hate type has a higher frequency.

3.4.3.5 Text length

Regarding the length of the text messages, we found that there seems to be no difference between the messages with "Hate speech" and with "None" (Table 3.5). We should bear in mind that in both cases the minimum number of words per message is three because we force it in our sampling process.

Table 3.5: Text lenght statistics of the messages classified as "Hate speech" or "None".

	hate speech	**none**
minimum	3	3
maximum	36	31
median	16	17
mean	15.92	16.15

We also compared the distributions of the main classes in our dataset with beanplots (Figure 3.9). We chose this representation because a beanplot is an alternative to the boxplot, where the individual observations are shown as small horizontal lines, the estimated density of the distributions is visible, and the average is shown [K+08]. Also, longer individual lines represent more instances for a certain value.

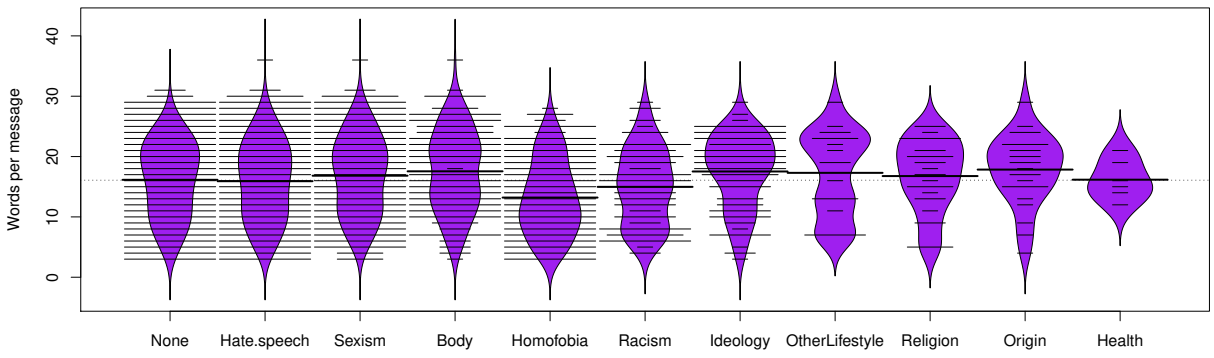

Figure 3.9: Beanplot of the number of words per message for each type of hate speech.

We found that the distributions between the messages with "Hate speech" and "None" are similar. However, some subtypes of hate speech have distinct distributions: "Homophobia" seems to have a lower number of words per message than the other classes; and some subtypes, namely "Other Lifestyle", "Religion", "Origin" and Health", have distributions with particular shapes. Regarding the differences in shape we should be cautious, because these classes have a lower number of instances, and the larger the number of instances the greater variance in the data.

3.4.3.6 Twitter metrics

We tried also to compare the different classes regarding the number of hashtags, mentions, retweets and URLs per message (Figure 3.10).

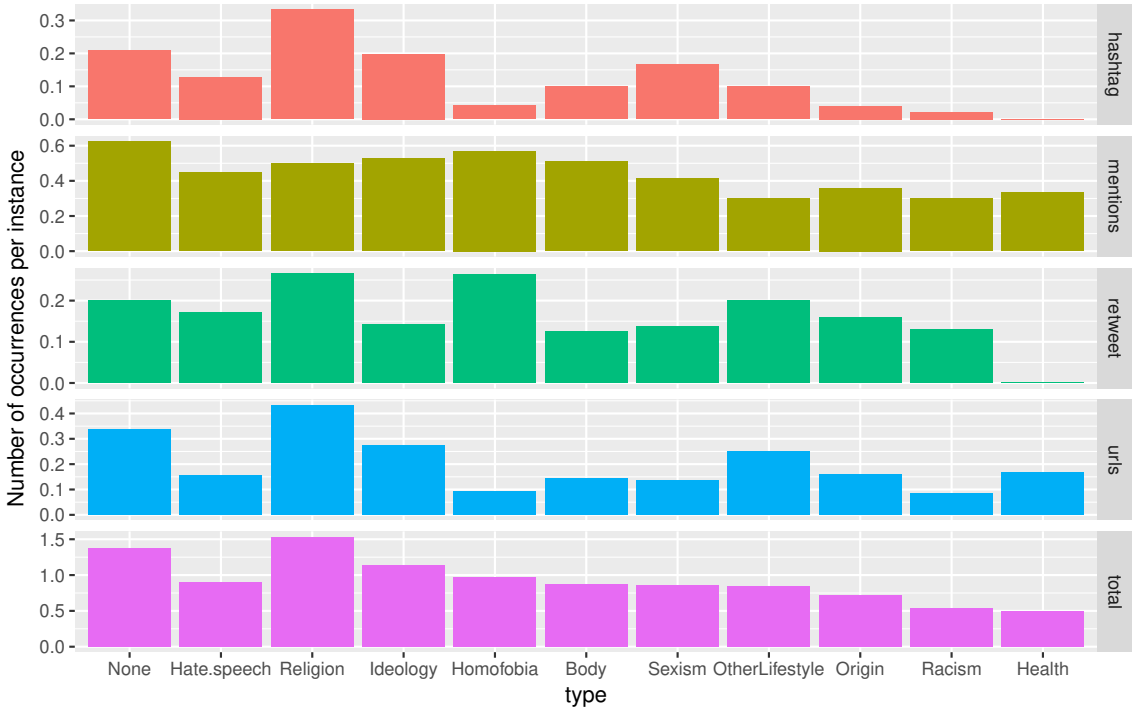

Figure 3.10: Number of hashtags, mentions, retweets, URLs per message for each type of hate speech and messages with "None".

We concluded that messages without hate speech ("None") tend to have more of these elements. We also noticed that hate speech based on "Religion" has more URLs and hashtags than the other hate types.

3.4.3.7 Vocabulary Size

Regarding the vocabulary size we aim to find information about the number of unique unigrams in the dataset. We consider unigrams the unique words that the dataset contains, excluding URLs, hashtags, retweet mark ("RT") and mentions. In the 5,668 messages of the dataset we found a total of 12,989 distinct unigrams, but in the messages with any type of "Hate speech" this number was 4,186 and in the messages with "None" 11,354. In the Figure 3.11, we try to understand if there is an influence of the total number of the messages in the number of distinct unigrams. In this figure we also compute the proportion between both using the equation PDU (proportion of distinct unigrams - Equation 3.1).

$$PDU(messages) = \frac{|getDistinctUnigrams(messages)|}{|messages|} \tag{3.1}$$

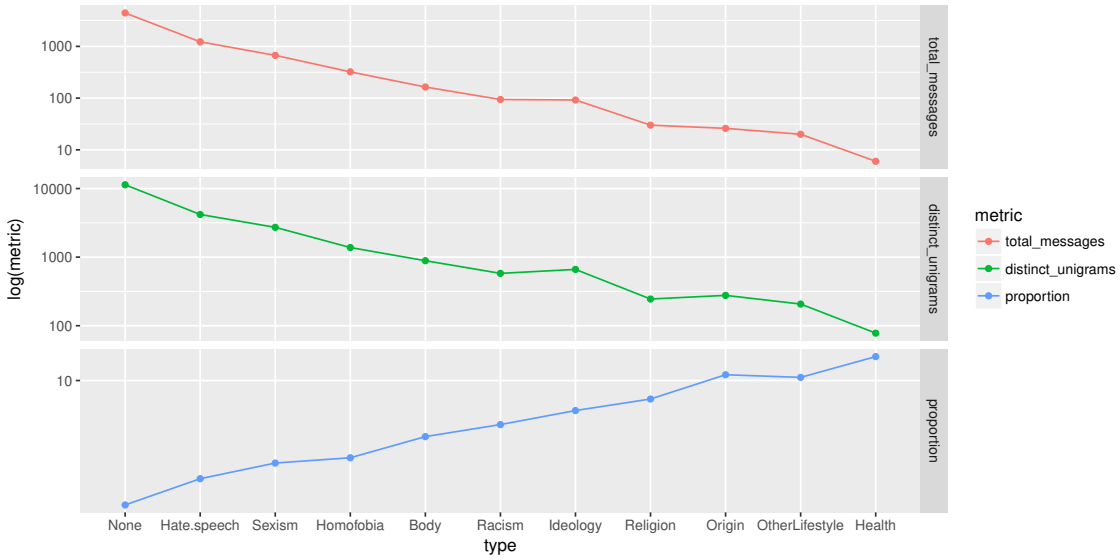

Figure 3.11: Number of distinct n-grams per message by hate type.

We found a general tendency that classes with more messages also have more distinct words. However this tendency does not apply when we compare “Religion” and “Origin”. In this case “Religion” has more messages but less distinct unigrams when compared to “Origin”. We also found that the more messages a class has the less distinct unigrams it has per message.

3.4.3.8 N-grams analysis

One way of analysing the content of text and frequencies of tokens in it is by using n-grams. With this procedure our goal was to discover if there are tokens that allow us to distinguish between messages with hate and messages without hate. We present here the method followed, the results and conclusions achieved.

Pre-processing To analyse the n-grams we preprocessed the text in the following way::

- Remove HTML.
- Remove URLs.
- Remove mention screen names.
- Remove RT (marker for retweet) from the beginning of the text.
- Replace all newlines with a whitespace.
- Replace multiple whitespaces with only one.
- Trim text after the previous steps.
- We extract n-grams, with N minor or equal to 10.

- In the n-grams with N equal to 1 we filter the stop words.

Tokens selection After the n-grams extraction we ended with a large number of columns as features. However much of this columns are sparce. This means that a lot of tokens are used in only a few number of messages. Sparsity can then be a problem in our analysis. First because of the needed time for processing, but mainly because we are not interested in keeping uncommon tokens that do not allow us to distinguish between classes. For instance, if we have only a few occurrences, we can not evaluate if the token is more common in the class "Hate speech" or "None".

For selecting the uncommon tokens, we then computed the frequency of each and we count the number of tokens that were having a certain frequency. Our goal was to find a threshold that would allow us to exclude uncommon tokens. We plot the total number of n-grams (y-axis) with a certain frequency (x-axis) (Figure 3.12).

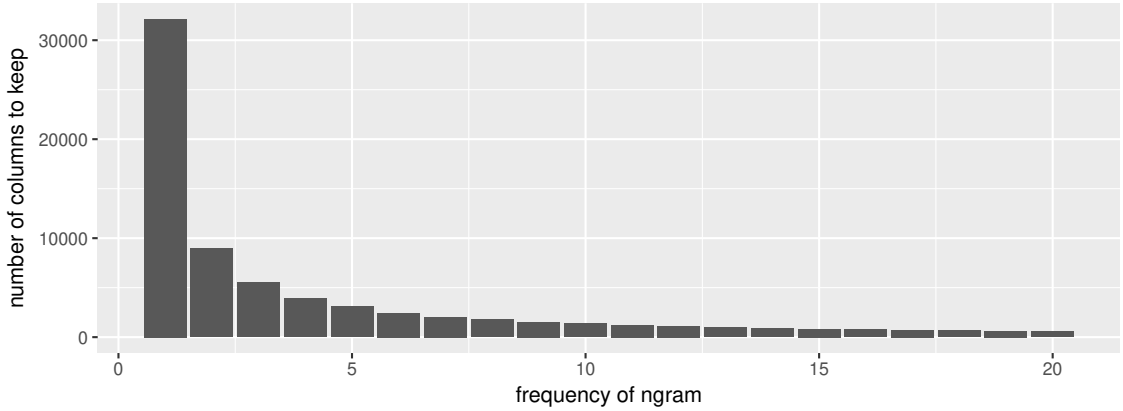

Figure 3.12: Types of hate frequencies in the dataset, plotted by frequency.

We concluded that the majority of the tokens had only one occurrence. Based on this plot we removed the tokens with less than 5 repetitions. In this case we are removing n-grams that are very unlikely in the context of the dataset and will not contribute to understand the differences between the several classes.

Selection of tokens in each class By using n-grams our goal was to understand if there are tokens more characteristic of each of the classes in our graph (presented in Appendix A). For that we developed an algorithm to get the top 10 more characteristic tokens for each class, based on a metric. This metric tries to privilege the tokens that are more frequent in the messages from the class, but at the same punish if they are also frequent in the messages not from the class (Equation 3.3).

Relative Frequency of a Token in a Set of Messages (RFTSM) Computes the relative frequency of a token n in a set of messages, by dividing the number of messages that has the token,

by the total number of messages (Equation 3.2). It works as an intermediate step for the final metric DFTPN.

$$RFTSM(messages, token_n) = \frac{|getMessagesWithToken(messages, token_n)|}{|messages|} \tag{3.2}$$

Difference between Frequencies of Token in the Positive and Negative messages from the class (DFTPN) Uses the RFTSM to compute the difference between the relative frequencies of a token in the messages that are from a class (positive messages) and the messages that are not from a class (negative messages) (Equation 3.3).

$$DFTPN(class_m, token_n) = RFTSM(positive(class_m), token_n) - RFTSM(negative(class_m), token_n) \tag{3.3}$$

For each of the classes, for each of the tokens, we computed this metric, then sort in descending order and got the tokens with the 10 highest values of DFTPN. As a result, for each class we got the lists of the more characteristic tokens that we present in the next section.

N-grams results In Table 3.6 we present a summary of the main results found. The results provided were translated manually to English. We selected the root class ("Hate speech") and the correspondent negative class (None) and the classes that are immediate children of hate speech ("Health", "Homophobia", "Ideology", "Origin", "Racism", "Religion", "Sexism" and "Other-lifestyle"). We present here classes with a representative number of instances and n-grams with only N equals to 1.

The main conclusion after extracting and analysing the tokens in each subtype of hate is that the words that are more common change depending on the hate type. This supports the idea that different types of hate have specific discourses [WH12b]. At the same time we can see that the words extracted in each class are slurs or stereotypes related with the topic that was assigned in the class name. This is more clear in "homophobia" (dyke, butch, dykes, pride, fagot, hetero, gay, gays), "ideology" (feminism, lefties, feminazi), "origin" (angola, latio, northeast, people, terrorism), "racism" (racism, white, black, nigger), "religion" (islam, muslim, mosque, bomb) and "sexism" (woman, fat - female, dumb - female, ugly - female, man).

We can also see that some tokens are shared between some of the classes. "Hate speech" shares (woman, fat, dumb, ugly and man) with the category "sexism". This is the case because we are using a hierarchical approach and one of the more common category in the dataset was "hate against women". Therefore in our approach "sexism" and "hate speech" will inherit and share a high number of instances against woman. On the other hand, "sexism" and "ideology" classes share also token (gorda). One explanation for this is that in the "ideology" class there are some words referring to feminism which is related with "sexism".

Table 3.6: Top-10 unigrams more common in the classes “Hate speech”, “None”, “Health”, “Homophobia”, “Ideology”, “Origin”, “Racism”, “Religion”, “Sexism” and “Other-lifestyle”. The original results in Portuguese can be found in Appendix D.

none	**hate speech**	**health**	**homophobia**	**ideology**
white	woman	music	dyke	feminist
pnr	fat (female)	listen	butch	feminists
angola	dumb (female)	owners	dykes	lefties
about	ugly (female)	would like	which	feminazi
bolsonaro	is	disease	pride	to be
left	dyke	who	fagot	feminism
trump	to be	dumb	hetero	to
all	butch	many	world	fat (female)
big	which	fuck	gay	is
now	man	guys	gays	father
origin	**racism**	**religion**	**sexism**	**other-lifestyle**
angola	racism	islam	woman	robber
latin	white	muslims	fat (female)	is
northeast	black	muslim	dumb (female)	good
enter	blacks	X5	ugly (female)	dead
people	is	europa	is	criminals
therefore	can	here	to be	laugh
can	nigger	safe (female)	man	can
light	because	bomb	which	people
terrorist	exists	turn	day	to
can	nothing	mosque	to	to date

3.5 Part-of-speech analysis

Other way of analysing the content of text and the relationships between words is using part-of-speech tagging.

3.5.1 POS Procedure

Using the packages NLP [Bal12], openNLP [Bal12] and openNLPmodels.pt [Bal12] we transform the text of the tweets and respective words. With the POS we replace the original word in the text by the role that the word has in the sentence. The most common roles are presented in this table. For this conversion we used the model pt-pos-maxent.bin provided in the package openNLPmodels.pt.

3.5.1.1 Classes differentiation with POS and n-grams

After transforming the text in POS tags we also used the same approach with n-grams, already described in the respective section (Section 3.4.3.8). We tried to understand which POS sequences are more frequent in each class. We present here the most frequent POS with length one (Table 3.7). The presented tags in this table are in Portuguese, because the translation to English is not straightforward.

Table 3.7: Top-10 POS more common in the classes "Hate speech", "None", "Health", "Homophobia", "Ideology", "Origin", "Racism", "Religion", "Sexism" and "Other-lifestyle".

hate speech	**health**	**homophobia**	**ideology**	**origin**	**racism**	**religion**	**sexism**	**other-lifestyle**
n	pronpers	prop	punc	n	adj	punc	n	n
prop	punc	punc	prop	punc	prop	art	vfin	vinf
vfin	n	art	adj	vfin	vpcp	vfin	adj	adv
pronpers	adv	prp	vfin	vinf	vfin	adj	pronpers	prp
adj	conjc	vpcp	prp	prop	n	adv	conjc	pronindp
adv	vfin	vinf	adv	prp	prondet	prop	prop	vpcp
conjc	prp	pronpers	vinf	adv	vinf	pronpers	vinf	art
punc	prondet	conjc	n	adj	punc	conjs	adv	conjc
vpcp	prop	adv	vpcp	pronpers	art	pronindp	conjs	punc
art	pronindp	adj	conjc	pronindp	adv	vpcp	vger	num

3.6 Annotation conclusions

One goal of this thesis was to annotate a dataset for Portuguese, because it is an important resource to promote research in hate speech detection in this language. More than this, the majority of previous studies was not offering datasets, and from the three existent datasets available, there is no main one defined for hate speech automatic detection in text.
We accomplished this main goal and we annotated a dataset that can be briefly described as in Table 3.8.

Table 3.8: Collected dataset in Portuguese Summary

metric	**dataset value**
Number of messages	5,668
Distinct users	1156
Number of annotators	A sample of 500 messages had 2 annotators.
Annotators agreement	K = 0.72
Classes	85 distinct classes. The main subtypes of hate speech used are Health, Homophobia, Ideology, Origin, Racism, Religion, Sexism, Other lifestyle
Instructions	Appendix C
Link	https://rdm.inesctec.pt/dataset/cs-2017-008

Additionally, our annotation method is innovative and specific for hate speech annotation. We use hierarchical labels and a DAG of classes, which is a structure that can integrate better the complexity of hate speech subtypes and its intersections.

Regarding the annotators agreement, we concluded that using the hierarchical approach allowed to achieve a better score. However the agreement between annotators is still an issue in identifying hate speech. We think that the complexity and subjectivity of the task makes the learning process not so immediate and more training is demanded for annotators. Therefore, in the communities aiming to reduce hate speech (e.g. EU Comission) it is necessary to assure that definitions of the concept are clear, providing rules, examples and more standardized guidelines. Besides, we also found that for different types of hate speech the agreement between annotators

can have contrasting values, which points out that some specific types of hate speech can be more difficult to identify than others. We also concluded that in the different studies diverse measures for agreement evaluation were used which makes more difficult to compare results.

We observed that from the total 5,668 messages, around 22% contains some type of hate speech, which is a better score than what was found in previous literature. We achieved this using two distinct methods. First, we look for specific pages that are conceived with the purpose of propagate hate, but after we also searched for specific hate related expressions.

An analysis on the dataset pointed out that the several types of hate speech present different patterns. First they have different number of messages in our sample and have distinct time occurrences. Regarding the number of words per messages, instances with "Hate speech" and "None" have similar distributions of the values, however messages with "Homophobia" seem to have less words. On the other hand, we also found differences regarding the number of hashtags, mentions, retweets and URLs per message. For the vocabulary size we concluded that classes with more messages have more distinct unigrams, however this is not the case when we compare the particular subtypes "Religion" and "Origin". Using n-grams we concluded also that the words that are more common, change depending on the hate type. Besides, we also added POS tags to the text and verified the same pattern, that the most common POS tags in each of the classes are not the same.

Finally, we developed in this chapter a new structure for annotating hate speech that uses hierarchical classes and we want to investigate if this methodology can improve the results in the classification tasks. That is the goal of our next chapter.

Chapter 4

Comparison of models using the annotated hate speech dataset

In the previous chapter we presented our methodology for annotating a dataset for hate speech classification in Portuguese. We used a hierarchical structure with classes represented as a directed acyclic category graph, where each class can have more than one parent. The advantage of using this structure is that it can better integrate nuances of the hate speech concept, such as the different subtypes of hate that exist and the intersectional relationship between them. Using a hierarchical structure for hate speech classification is an innovative method and therefore in the present section we want to investigate the impact that the used categories can have in detecting hate speech online. One first consequence is immediate. Using this complex structure of classes allows for a multilabel classification on hate speech and therefore to train and detect for different subtypes of hate speech.

Moreover, we want to investigate another possible advantage of this structure. We want to understand if the information of the subtypes of hate speech can be used in the prediction of the class hate speech in itself. This approach is based on the idea that different subtypes of hate have different discourses and demand distinct models [WH12b]. In order to accomplish this final goal of our thesis we conducted an experiment following a methodology with training, validation and test phases. We describe the followed procedure, results and conclusions achieved in this section.

4.1 Methodology

In the final task of our thesis, the goal is to compare different models and investigate the effect of using a hierarchical structure for classification in hate speech detection. For this we followed a methodology based on training, validation and testing [Bis06]. We conducted an experiment with two different variations in the training phase. While in Method 1, we train considering only the hate speech class, in the Method 2 we consider more labels, correspondent to each subtype of hate. This variation is based on the idea that different subtypes of hate speech have distinct

discourses [WH12b], for instance, anti-hispanic speech refers more border crossing, while anti-semitic speech often refers more to money. Then, in Method 2, we take advantage of the available labels for each subtype of hate speech and we try to improve the procedure of feature extraction with this labels information. Both methods 1 and 2 target to identify the more general hate speech class. The summary of the used pipeline is presented in the Fig. 4.1. We describe our methodology with more detail in the next subsections .

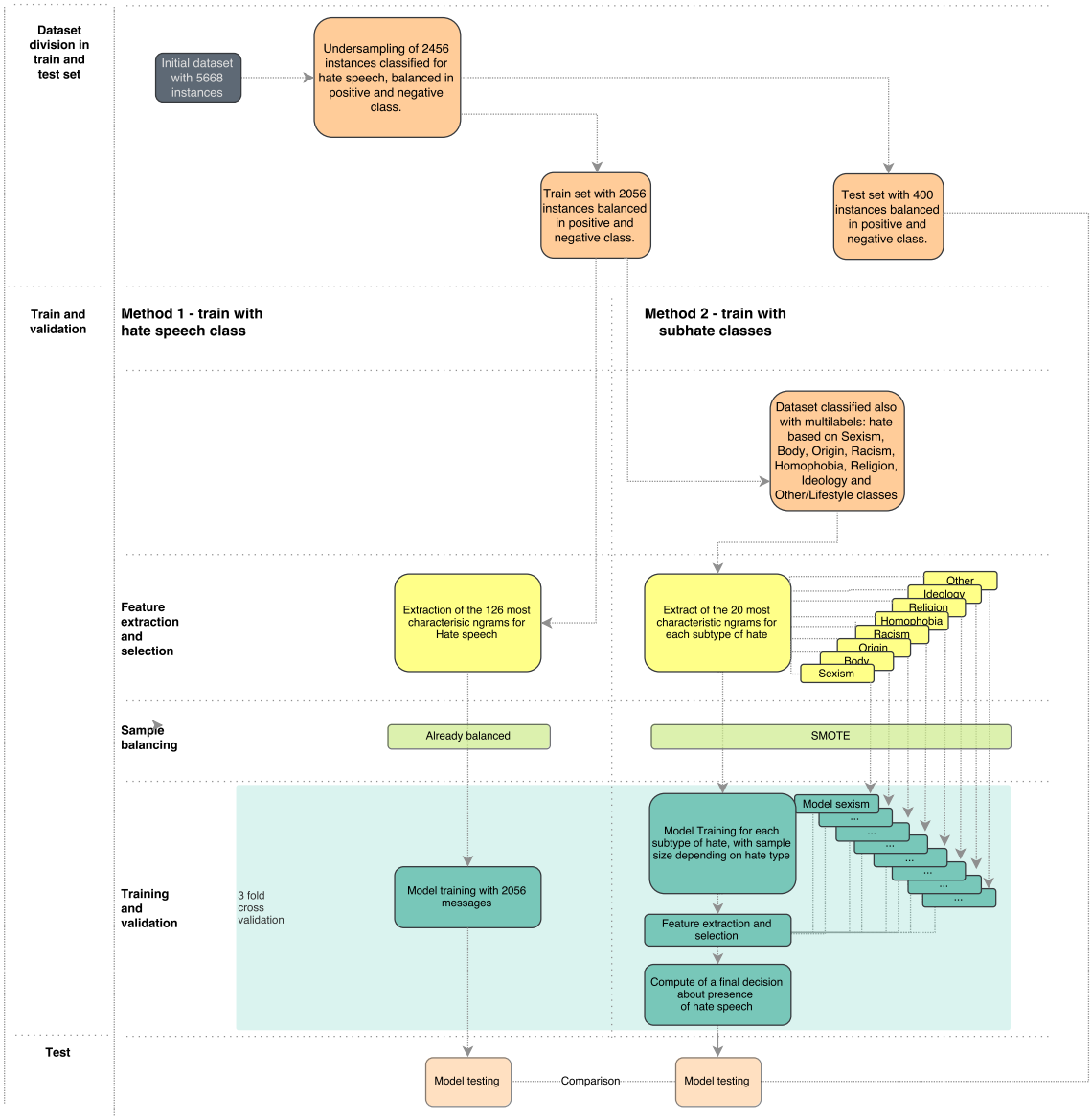

Figure 4.1: Pipeline used for model comparison using the Portuguese hate speech detection dataset.

4.1.1 Sampling and dataset division in train and test set

In order to conduct a procedure with training, validation and testing phases, it was necessary to split our data in train and test set. As a source of messages with hate speech, we started by using the dataset collected in Section 3, with a total of 5,668 messages. However, this dataset is unbalanced and contains only 1,228 messages with hate speech, opposing to 4,440 messages without hate speech. When applying machine learning classification algorithms, classes frequencies can have an impact in the performance of the model, because most learning systems usually assume that training sets are balanced [BPM04]. One way of preventing these situations is to use sampling techniques such as down-sampling or over-sampling. In the case of our experiment, we undersampled the messages without hate speech and ended with 2,456 instances, half classified as "hate speech", and half classified as "none".

After this first step, we randomly divided the instances in train (2,056) and test (400) sets, ensuring again the same proportion of negative and positive class instances in both sets. Regarding the distinct subtypes of hate, in the final data sets we end with distinct frequencies from the considered classes (Table 4.1).

Table 4.1: Positive class frequencies in train and test sets used in the conducted experiment.

Class	Train set	Test set
Hate speech	1028	200
Sexism	563	109
Body	144	20
Origin	24	2
OtherLifestyle	17	3
Racism	73	21
Homophobia	268	54
Religion	26	4
Ideology	79	13

4.1.2 Train and validation phases

For training and validating we applied two different methods that we want to compare. These methods were distinct in the feature extraction, sample balancing and training.

Feature extraction and selection Regarding the features, we chose n-grams because this method is used widely in the problem of hate speech detection [BW16, NTT+16, WH16, LF14, BYH+, GS04, DWMW17, BGGV17], and it is presented as a base that provides good results [SW17] on top of which other features can be added. After extracting the n-grams we got 7,440 features in our dataset, which is a much larger number comparing to the 2,056 instances. When a classification task has a large number of dimensions, the problem of curse of dimensionality can arise [Bis06]. This metaphor is used due to the disadvantages of using too many features such as: the complexity of the problem raises exponentially, which increases the time necessary for computing

the models [Bis06]; it can lead to lack of data [Bis06]; and lack of model identifiability, overfitting and numerical instabilities [VF05].

We decided then to conduct a procedure for feature selection in order to reduce the number of features. Based on the metric RFTSM and DFTPN presented in Subsection 3.4.3.8, we selected the 20 most representative n-grams for each of the classes. However, in this experiment we are comparing two methods: the first considers only the hate speech class, while the second considers more subtypes of hate. This difference implies that in the first case we have only 20 distinct n-grams, while in the second case there are the 20 more distinct n-grams for each of the classes: Sexism, Body, Origin, Other/Lifestyle, Racism, Homophobia, Religion and Ideology. For this reason, in order to make both methods more comparable, we computed the number of distinct n-grams used in the second method, and assured that we are using the same number in the first. We end up using as features the 126 more characteristic n-grams of the class hate speech, in the method 1; while we were using the 20 more characteristic n-grams for each of the 8 subtypes of hate speech considered, for method 2.

Sample balancing We already presented an initial sampling step conducted, commonly for method 1 and 2, which consisted in balance the dataset for the hate speech class. Therefore the conducted procedure assures that in the Method 1, where the train considers only the hate speech class, no extra balancing is needed, because the number of messages from the positive class is already the same as the negative class. However, for Method 2, the train regards other subclasses, that are unbalanced (Table 4.1). In this case we have a problem that is related not only with unbalanced classes, but also that some classes are much more infrequent. In this case we can use SMOTE [TT13], both for replicating instances from the positive class and undersampling the negative class. When using SMOTE we aimed to achieve a proportion of around 25% of positive class to 75% of the negative class in a total of 400 messages. Because SMOTE generates simulated data, and that can bring error to the model, we applied it only in the classes with less than 100 instances (Origin, Other-Lifestyle, Racism, Religion and Ideology). For the other subtype classes (Sexism, Body and Homophobia) we applied undersampling of the majority class to reach a proportion of 50% between positive and negative class.

Training For learning the classification of hate speech messages, different algorithms were used. We chose the algorithms in order to use different types of models: linear (logistic regression and support vector machines linear); neural networks (multilayer perceptron); trees (recursive partitioning); ensemble methods (random forest, Xgboost). For the application of these algorithms, we used the package caret in R [Kuh08] and the function train. This function allows to specify a huge amount of learning parameters. We defined that the optimization should be conducted for the ROC metric; regarding the tune length of the hyper parameters, we specified a length 3.

In the case of Method 1, training only with hate speech class, each of the algorithms was used to train the model. In the case of the Method 2, training with the subtype of hate speech classes, first we predict if each message contains a subtype of hate speech. After, the result of

these predictions, is used as features for predicting hate speech. We considered that messages contain hate speech when they contain at least one subtype of hate speech.

Validation Regarding the validation procedure, we used the options available also in the caret library [Kuh08]. In this case, the train function can receive a trainControl structure, with parameters. We specified the method cross validation, with three folds. We chose a lower number for the number of folds due to the low number of instances in some classes. Literature in the topic points out that, if the validation set is small, it will give relatively noisy estimate of predictive performance [BPM04].

4.1.3 Test

For the testing we also used the package caret [Kuh08]. The respective function predictedClasses was used to apply the trained models to the test set, while the function confusionMatrix was used to compare the real classes of the test set with the predicted classes by the models. The function confusionMatrix provides different measures: true positives, false positives, true negatives, false negatives and accuracy. With this values we also computed Precision, Recall and F1 [Pow11]. As a summary we present the main differences between the two methods that we aim to compare (Table 4.2).

Table 4.2: Main differences between the two compared methods.

	Method 1	**Method 2**
Classes considered in the processing	Hate speech	Sexism, Body, Origin, Other/Lifestyle, Racism, Homophobia, Religion and Ideology
Balancing of samples	Balance in the messages with and without Hate speech.	SMOTE was used. The majority classes was down-sampled and the minority classes was over-sampled.
Features	The 126 more characteristic n-grams of the class Hate speech.	The 20 more distinct n-grams for each of the subtype of hate speech classes.
Training	Train only with hate speech class.	Train with the subtype of hate classes. After predicting each subtype of hate, a final class of hate speech was computed.

In the next section we present the results of the performance for both models in the final test set with 400 instances.

4.2 Results and discussion

Regarding the comparison between both methods we analysed the prediction results in a test set with 400 instances, never seen in the training phase of the learning process. We evaluated the confusion matrix frequencies. In this context, presence of hate speech is our positive class, and the absence of hate speech is our negative class. We use for Method 1 the abbreviation “unimodel” and for the Method 2 the abbreviation “multimodel”.

We concluded (Table 4.3) that both methods identify an higher number of true positives when using the MLP algorithm, achieving 144 positive instances well classified in the unimodel and 153 in the multimodel. Regarding the true negatives, both methods identified an higher number using the Rpart algorithm, achieving 195 positive instances well classified in the unimodel and 189 in the multimodel.

Table 4.3: Confusion matrix metrics summarized for both methods and different algorithms.

	TP		**FP**		**TN**		**FN**	
algorithm	unimodel	multimodel	unimodel	multimodel	unimodel	multimodel	unimodel	multimodel
LogReg	110	134	12	37	188	163	90	66
MLP	144	153	35	41	165	159	56	47
SVMLinear	141	147	28	42	172	158	59	53
RF	131	141	20	33	180	167	69	59
Xgboost	119	144	15	35	185	165	81	56
Rpart	57	83	5	11	195	189	143	117

We also noticed (Figure 4.2) that the multimodel has a better frequency of hate speech class well classified (TP), while the unimodel performs better in identifying the absence of hate speech (TN). However at the same time, the multimodel has more instances classified wrongly as hate speech, while the unimodel has more instances classified wrongly as absence of hate speech. This denotes a tendency that the unimodel uses more the label absence of hate speech, while the multimodel uses more the label hate speech, and therefore each of the models identify correctly, but also misclassify more, in the respective class.

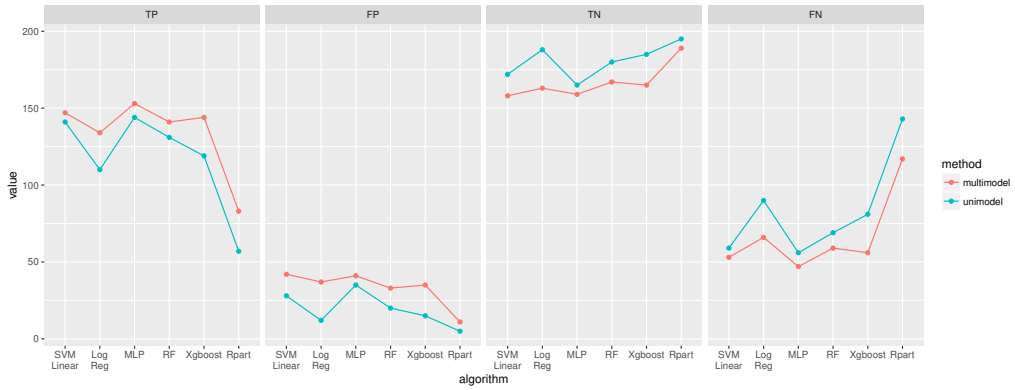

Figure 4.2: Graphical comparison of the confusion matrix metrics between unimodel and multimodel.

In order to understand better how this tendency affects the performance of the models we computed other metrics, such as Accuracy, Precision, Recall, difference between Precision and Recall and F1. We concluded (Table 4.4) that the best Accuracy was achieved in the unimodel using SVMLinear (0.778), and for the multimodel using the MLP (0.78); the best Precision was achieved using Rpart both for unimodel (0.778), and for multimodel (0.883); the best Recall was achieved using SVMLinear both for the unimodel (0.720), and for multimodel (0.765); and the

best F1 was achieved for the unimodel using SVMLinear (0.764), and for the multimodel using the MLP (0.777).

Table 4.4: Performance metrics summarized for both methods and different algorithms

	Accuracy		Recall		Precision		F1	
algorithm	unimodel	unimodel	unimodel	unimodel	unimodel	unimodel	unimodel	unimodel
LogReg	0.740	0.743	0.550	0.670	0.902	0.784	0.683	0.722
MLP	0.773	0.780	0.720	0.765	0.804	0.789	0.760	0.777
SVMLinear	0.783	0.763	0.705	0.735	0.834	0.778	0.764	0.756
RF	0.778	0.770	0.655	0.705	0.868	0.810	0.746	0.754
Xgboost	0.760	0.773	0.595	0.720	0.888	0.804	0.713	0.760
Rpart	0.630	0.680	0.285	0.415	0.919	0.883	0.435	0.565

We conclude (Figure 4.3) that regarding the accuracy it is difficult to distinguish the performance of the algorithms when the two methods were followed. However we find differences in the values for Precision and Recall. The unimodel has a better Precision, which means that it has a better proportion between positive instances correctly identified and the total positive instances retrieved. This is the case, because the multimodel has more false positives (FP, Figure 4.2). On the other hand, the multimodel has a better Recall which means that a larger proportion of hate speech messages is retrieved, over the total existent hate speech messages.

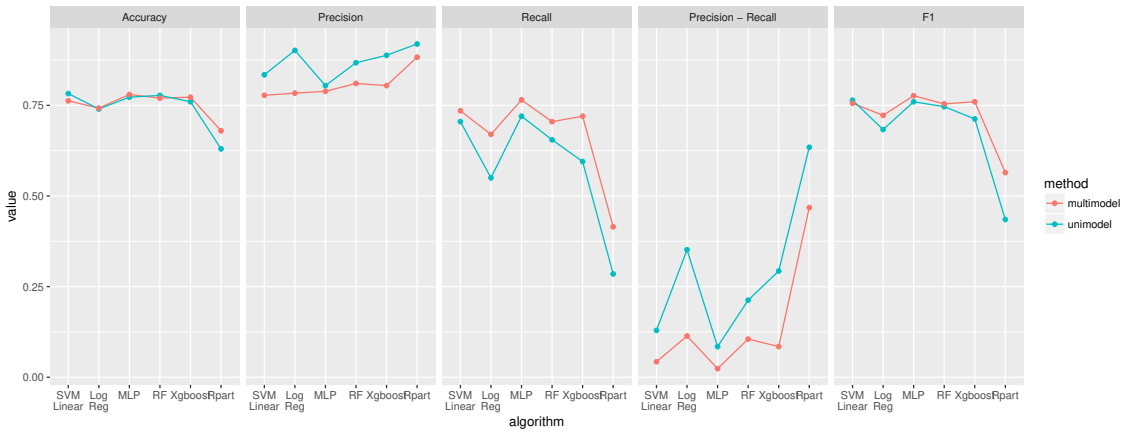

Figure 4.3: Graphical comparison of metrics, between unimodel and multimodel.

We can also notice that, regarding the difference between Precision and Recall, the unimodel has a larger difference between these metrics. In this method, if on one hand we have a very good precision, on the other hand we have a worse recall, while in the multimodel both values are more similar. This points out that both categories are identified with a more similar performance in the multimodel, and the error occurs in a more similar way in both positive and negative classes. Finally, we found that the measure of F1 is slightly better in the multimodel (except when using SVM linear algorithm).

4.3 Conclusions of the comparison between the two models

In this Chapter we aimed to investigate if the information of the subtypes of hate speech could be used to improve the prediction of the class hate speech in itself. We followed a pipeline as a guide to our work and we noticed some challenges in this task. First, if on one hand hate speech messages are already rare, when we consider the subtypes of hate speech we are in the presence of even more rare occurrences. Also, the distinct subtypes of hate speech classes appear in distinct frequencies. To overcome this problem in our experiment we used SMOTE. On the other hand, because we aimed to prove that using a hierarchical structure of classes can help to improve results, we used n-grams with lenght one, that are simplistic and non specific to hate speech features. In a first moment this lead us to an exaggerated number of features, when compared to the available instances. We then selected the features more characteristic of our positive classes.

Based on the results obtained, we observed that the performance of the multimodel seems to be slightly better than the unimodel, because it seems to have better results in the F1 metric. Besides, our method helps to identify a larger number of hate speech messages. This is the case because it has a better recall. However, at the same time the multimodel has a worse precision. Usually this is a trade off when comparing models. Models with better precision are more reliable when identifying the positive class, while models with a better recall are able to identify a higher number of instances in the positive class. In the case of hate speech detection, we think that it is better to privilege a model with better recall, because we want to identify the highest possible number of these messages. Even when messages are classified wrongly as hate speech, another layer of evaluation can be used to better filter hate speech messages. We should also point out that using the subtypes of hate speech provides us with important information regarding the description of online hate speech. Using the multimodel allows us to better understand the targets of hate speech, which is a valuable information when tackling this problem.

Finally, this experiment can be extended in the future, in order to improve results. One first priority can be to increase the number of instances from the classes low represented. With a low number of instances models perform worse, and in our approach we used SMOTE algorithm to solve this problem, which simulates new data and can introduces error in our model. Additionally, regarding the used features, more complex approaches can be tested in the future, such as with particular and related with the hate speech features (Figure 2.9). In this case features like the othering discourse should be considered, however focusing on each subtype of hate.

Chapter 5

Conclusions and Future work

In this chapter we analyse the work conducted in this thesis in terms of our goals, how we accomplish them and the obtained results. After that, we regard the open questions and possible future work in the area.

5.1 Goals of the work

With our thesis we aimed to enrich the field of automatic hate speech detection in text. More in particular we had three main goals. The first was to understand what has been done so far in this field. To accomplishing this we conducted an overview of the topic. We concluded that hate speech has been defined in several platforms, from social networks to other organizations. We critically summarized these perspectives and proposed a single definition of hate speech to guide our work. We also complemented our definition with rules and examples and compared it with other related concepts, such as hate, cyberbullying, abusive language, discrimination, profanity, toxicity and flaming.

To better understand and describe the state of art on this topic we conducted a systematic literature review. We found that the number of studies and papers published in automatic hate speech detection in text is limited and researchers tend to start by collecting and classifying new messages, but these datasets are, in the majority of the cases, not made available. This practice slows down the progress in this research field, because it is necessary to share data in order to compare approaches and results. Nevertheless, we found three available datasets, in English and German. Regarding the features used in the collected documents, we found that general text mining approaches are used (e.g. n-grams, POS, rule based approaches, sentiment analysis, deep learning features such as word2vec). Complementary, specific hate speech detection features are used as well (e.g. "othering" language, superiority of the ingroup, focus on stereotypes).

From this overview in the topic we found that no research had been conducted for Portuguese so far. We then proposed a second goal for our work. We aimed to collect a dataset for Portuguese

and make it publicly available. This is a relevant contribution, not only because this language is one of the most spoken in the world, but also because we used an innovative structure for hate speech classification. We considered that the best structure for this problem are hierarchical labels, represented as a DAG of classes. This is the structure that can integrate the complexity of hate speech subtypes and its intersections.

In order to obtain our data we crawled Twitter for messages and manually annotated them following a set of rules, that are also a valuable product of our work. We accomplish our goal and we annotated a dataset with 5,668 messages from 1,156 distinct users, where 85 distinct classes of hate speech were considered. From the total 5,668 messages, around 22% contained some type of hate speech. With this percentage, we improved the ratio of hate speech messages, in comparison to what was found previously in other studies, saving resources in the annotation procedure. We achieved this using two distinct methods: first, we looked for specific pages that are conceived with the purpose of propagate hate; and after we searched for specific hate related expressions, as well.

We also described the collected dataset. Regarding the annotators agreement, we concluded that using the hierarchical approach allowed us to achieve a better score. However the agreement between annotators is still an issue in identifying hate speech. Other analysis pointed out that the several types of hate speech present different characteristics: the number of messages is distinct; the time occurrences (hour and weekday); vocabulary size (classes with more messages have more distinct unigrams); most common words are different (n-grams); and most common POS are different as well.

A final goal of our thesis was to investigate the potential advantages of using hierarchical classes to annotate a dataset in hate speech automatic detection. One first consequence of using a hierarchical structure for hate speech classification is immediate. Using this complex structure of classes allows for a multilabel classification on hate speech and therefore to train and detect for different subtypes of hate speech. However, we wanted to understand if the information of the subtypes of hate speech can be used in the prediction of the class hate speech in itself. For this, we used the dataset annotated for Portuguese and we conducted an experiment with training, validation and test phases. In this experiment we used two different conditions: we called unimodel to the model that was using only the information of the hate speech class; and multimodel to the model that was using the information of the classes Sexism, Body, Origin, Other/Lifestyle, Racism, Homophobia, Religion and Ideology.

As conclusions of our experiment, we noticed some challenges in this task. First, if on one hand hate speech messages are already rare, when we considered the subtypes of hate speech, we are in the presence of even more rare occurrences. To overcome this problem we used SMOTE, a method for instances replication towards a balancing dataset. On the other hand, because we aimed to prove that using a hierarchical structure of classes can help to improve results, we used very simplistic features and non specific to hate speech. In a first moment this lead us to an exaggerated number of feature, thus we selected the features more characteristic of our positive classes.

The performance of the multimodel seemed to be slightly better than the unimodel in the F1

metric. Besides, our method helps to identify a larger number of hate speech messages. This is the case because it has a better recall, in detriment of the precision. In the case of hate speech detection, we think that it is better to privilege recall, because we want to identify the highest possible number of these messages. To improve our model, another layer of evaluation can be used to better filter false positive hate speech messages.

5.2 Future work

Finally during the conducted research we also spotted some opportunities in the field. From our systematic literature review we found a lack of open source platforms that automatically classify hate speech; no comparative studies that would summarize the approaches conducted so far; and, because the majority of the research was conducted only in English, languages such as French, Mandarin, Portuguese or Spanish had no advances in this area.

Regarding the findings from our dataset annotation procedure, we face troubles in the agreement between annotators. We think that the complexity and subjectivity of the task makes the learning process not so immediate and more training is demanded from the annotators. Therefore, in the communities aiming to reduce hate speech (e.g. EU Comission) it is necessary to assure that definitions of the concept are clear, providing rules, examples and more standardized guidelines.

Finally, also the conducted experiment can be extended in the future, in order to improve results. One first priority can be to increase the number of instances from the classes with low representation. With a low number of instances models perform worse, and in our approach we used SMOTE algorithm to solve this problem, which simulates new data but can introduce error in our model. Additionally, regarding the used features, we only used n-grams with length one, and more complex approaches can be tested in the future. Also features more particular and related with the hate speech concept can provide better solutions.

Finally, with our thesis we tried to identify the main problems in the field of automatic hate speech in text and to tackle some. Undoubtedly this is a field that still needs more research and with unlimited challenges.

References

[20117] TA-COS 2016. Ta-cos 2016. Available in http://www.ta-cos.org/, accessed last time in February 2017, 2017.

[ACL17] ACL. Alw1: 1st workshop on abusive language online. Available in https://sites.google.com/site/abusivelanguageworkshop2017/home, accessed last time in February 2017, 2017.

[AS17] Swati Agarwal and Ashish Sureka. Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr micro-blogging website. *arXiv preprint arXiv:1701.04931*, 2017.

[Bal12] Jason Baldridge. The opennlp project. Available in http://opennlp.apache.org/index.html , accessed last time in May 2017, 2012.

[BGGV17] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee, 2017.

[Bis06] Christopher M. Bishop. Pattern recognition. *Machine Learning*, 128:1–58, 2006.

[BNP+14] Jamie Bartlett, Richard Norrie, Sofia Patel, Rebekka Rumpel, and Simon Wibberley. Misogyny on twitter. *DEMOS*, 05:18, 2014.

[BPM04] Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter*, 6(1):20–29, 2004.

[BW14] Peter Burnap and Matthew L. Williams. Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making. In *Proceedings of IPP*, pages 1–18, 2014.

[BW15] Pete Burnap and Matthew L. Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242, 2015.

[BW16] Pete Burnap and Matthew L. Williams. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science*, 5(1):11, 2016.

[BYH+] Tanvi Banerjee, Amir H. Yazdavar, Andrew Hampton, Hemant Purohit, Valerie L. Shalin, and Amit P. Sheth. Identifying pragmatic functions in social media indicative of gender-based violence beliefs. *Manuscript submitted for publication*.

REFERENCES

[CH15] Keith Cortis and Siegfried Handschuh. Analysis of cyberbullying tweets in trending world events. In *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business*, page 7. ACM, 2015.

[Che11] Ying Chen. *Detecting Offensive Language in Social Medias for Protection of Adolescent Online Safety*. PhD thesis, The Pennsylvania State University, 2011.

[CHI17] CHI2017. 2017 workshop on online harassment. Available in http://social.umd.edu/woh/, accessed last time in February 2017, 2017.

[CLi16] CLiPS. Hades. Available in https://github.com/clips/hades , accessed last time in May 2017, 2016.

[CON17] CONTACT. Interdisciplinary conference on hate speech. definitions, interpretations and practices. Available in https://sites.google.com/site/abusivelanguageworkshop2017/home, accessed last time in May 2017, 2017.

[Cro17] CrowdFlower. Data for everyone. Available in https://www.crowdflower.com/data-for-everyone/ , accessed last time in May 2017, 2017.

[DAKAA15] Ali A. Dashti, Ali A. Al-Kandari, and Hamed H. Al-Abdullah. The influence of sectarian and tribal discourse in newspapers readers' online comments about freedom of expression, censorship and national unity in kuwait. *Telematics and Informatics*, 32(2):245–253, 2015.

[Dav17] Thomas Davidson. Automated hate speech detection and the problem of offensive language. Available in https://github.com/t-davidson/hate-speech-and-offensive-language , accessed last time in May 2017, 2017.

[DC00] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 256–263. ACM, 2000.

[DdJOT12] Maral Dadvar, Franciska de Jong, Roeland Ordelman, and Dolf Trieschnigg. Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop*, pages 23–25. University of Ghent, 2012.

[Dic17] Cambridge Dictionary. Profanity. Available in https://dictionary.cambridge.org/dictionary/english/profanity , accessed last time in June 2017, 2017.

[DMEY16] Sara Douglass, Sheena Mirpuri, Devin English, and Tiffany Yip. They were just making jokes: Ethnic/racial teasing and discrimination among adolescents. *Cultural Diversity and Ethnic Minority Psychology*, 22(1):69, 2016.

[DMM08] Marie-Catherine De Marneffe and Christopher D Manning. Stanford typed dependencies manual. Technical report, Technical report, Stanford University, 2008.

[DRL11] Karthik Dinakar, Roi Reichart, and Henry Lieberman. Modeling the detection of textual cyberbullying. *The Social Mobile Web*, 11(02), 2011.

REFERENCES

[DVCD+17] Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity*, pages 86–95, 2017.

[DWMW17] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*, 2017.

[DZM+15] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, pages 29–30. ACM2, 2015.

[Eys16] Eyspahn. Online hate speech modeling using python and reddit comment data. Available in https://github.com/eyspahn/OnlineHateSpeech_PyLadiesSea , accessed last time in May 2017, 2016.

[Fac13] Facebook. What does facebook consider to be hate speech? https://www.facebook.com/help/135402139904490, accessed last time in February 2017, 2013.

[FBI15] FBI. 2015 hate crime statistics. Available in https://ucr.fbi.gov/hate-crime/, accessed last time in February 2017, 2015.

[Fox13] Zoe Fox. Top 10 most popular languages on twitter. Available in http://mashable.com/2013/12/17/twitter-popular-languages/, accessed last time in May 2017, 2013.

[GHH07] Radhouane Guermazi, Mohamed Hammami, and Abdelmajid Ben Hamadou. Using a semi-automatic keyword dictionary for improving violent web site filtering. In *Signal-Image Technologies and Internet-Based System, 2007. SITIS'07. Third International IEEE Conference on*, pages 337–344. IEEE, 2007.

[GL15] Fabio Giblietto and Yenn Lee. To be or not to be charlie: Twitter hashtags as a discourse and counter-discourse in the aftermath of the 2015 charlie hebdo shooting in france. In *4th Workshop on Making Sense of Microposts (# Microposts2014)*, 2015.

[GLG+12] Matthias Gamer, Jim Lemon, Maintainer Matthias Gamer, A Robinson, and W Kendall's. Package 'irr'. *Various coefficients of interrater reliability and agreement*, 2012.

[Gre04] Edel Greevy. *Automatic text categorisation of racist webpages*. PhD thesis, Dublin City University, 2004.

[Gro17] Stanford NLP Group. The stanford nlp group. Available in http://nlp.stanford.edu/, accessed last time in February 2017, 2017.

[GS04] Edel Greevy and Alan F. Smeaton. Classifying racist texts using a support vector machine. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 468–469. ACM, 2004.

REFERENCES

[GZDL15] Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230, 2015.

[Hat17] Reporting Hate. Hate speech conference i.h.d.i.p. Available in http://reportinghate.eu/contact2017/, accessed last time in May 2017, 2017.

[HCT07] Pei-Yi Hao, Jung-Hsien Chiang, and Yi-Kun Tu. Hierarchically svm classification based on support vector clustering method and its application to document categorization. *Expert Systems with applications*, 33(3):627–635, 2007.

[Her16] Alex Hern. Facebook, youtube, twitter and microsoft sign eu hate speech code. Available in https://www.theguardian.com/technology/2016/may/31/facebook-youtube-twitter-microsoft-eu-hate-speech-code, accessed last time in February 2017, 2016.

[HL14] Yannis Haralambous and Philippe Lenca. Text classification using association rules, dependency pruning and hyperonymization. *arXiv preprint arXiv:1407.7357*, 2014.

[HTB16] Sarah Hewitt, Thanassis Tiropanis, and Christian Bokhove. The problem of identifying misogynist language on twitter (and other online social spaces). In *Proceedings of the 8th ACM Conference on Web Science*, pages 333–335. ACM, 2016.

[ILG16] ILGA. Hate crime and hate speech. Available in http://www.ilga-europe.org/what-we-do/our-advocacy-work/hate-crime-hate-speech, accessed last time in October 2016, 2016.

[Jig17] Jigsaw. Perspective api. Available in https://www.perspectiveapi.com/ , accessed last time in June 2017, 2017.

[K+08] Peter Kampstra et al. Beanplot: A boxplot alternative for visual comparison of distributions. *Journal of statistical software*, 28(1):1–9, 2008.

[Kag13] Kaggle. Detecting insults in social commentary. Available in https://www.kaggle.com/c/detecting-insults-in-social-commentary/data, accessed last time in February 2017, 2013.

[KG16] Till Krause and Hannes Grassegger. Facebook's secret rules of deletion. Available in http://international.sueddeutsche.de/post/154543271930/facebooks-secret-rules-of-deletion, accessed last time in February 2017, 2016.

[Kom16] Panos Kompatsiaris. Whitewashing the nation: racist jokes and the construction of the african 'other'in greek popular cinema. *Social Identities*, pages 1–16, 2016.

[Kri04] Klaus Krippendorff. *Content analysis: An introduction to its methodology*. Sage, 2004.

[Kuh08] Max Kuhn. Caret package. *Journal of Statistical Software*, 28(5):1–26, 2008.

[KvdE16] Giselinde Kuipers and Barbara van der Ent. The seriousness of ethnic jokes: Ethnic humor and social change in the netherlands, 1995–2012. *humor*, 29(4):605–633, 2016.

REFERENCES

[KW13] Irene Kwok and Yuzhou Wang. Locate the hate: Detecting tweets against blacks. In *Association for the Advancement of Artificial Intelligence*, 2013.

[Lea15] Anti-Defamation League. The trap of masculinity: how sexism impacts boys and men. Available in https://www.adl.org/sites/default/files/documents/assets/pdf/education-outreach/trap-of-masculinity.pdf, accessed last time in May 2017, 2015.

[LF14] Shuhua Liu and Thomas Forss. Combining n-gram based similarity analysis with sentiment analysis in web content classification. In *International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, pages 530–537, 2014.

[LF15] Shuhua Liu and Thomas Forss. New classification models for detecting hate and violence web content. In *Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference on*, volume 1, pages 487–495. IEEE, 2015.

[Mal14] Wilson Jeffrey Maloba. *Use of regular expressions for multi-lingual detection of hate speech in Kenya*. PhD thesis, iLabAfrica, 2014.

[MT16] Yashar Mehdad and Joel Tetreault. Do characters abuse more than words? In *Proceedings of the SIGdial 2016 Conference: The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303, 2016.

[NS15] B. Nandhini and J. I. Sheeba. Cyberbullying detection and classification using information retrieval algorithm. In *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015)*, page 20. ACM, 2015.

[NTT+16] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153. International World Wide Web Conferences Steering Committee, 2016.

[OC14] Andre Oboler and Karen Connelly. Hate speech: A quality of service challenge. In *e-Learning, e-Management and e-Services (IC3e), 2014 IEEE Conference on*, pages 117–121. IEEE, 2014.

[oCU17] United Nations Alliance of Civilizations (UNAOC). #spreadnohate: A global dialogue on hate speech against migrants and refugees in the media. Available in https://www.unaoc.org/what-we-do/projects/hate-speech/, accessed last time in May 2017, 2017.

[Pow11] David Martin Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *International Journal of Machine Learning Technology*, 2(1):37–63, 2011.

[Red02] Vasu Reddy. Perverts and sodomites: Homophobia as hate speech in africa. *Southern African Linguistics and Applied Language Studies*, 20(3):163–175, 2002.

[RH16] Elaheh Raisi and Bert Huang. Cyberbullying identification using participant-vocabulary consistency. *arXiv preprint arXiv:1606.08084*, 2016.

REFERENCES

[RRC+17] Bjorn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*, 2017.

[SMC+16] Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. Analyzing the targets of hate in online social media. *arXiv preprint arXiv:1603.07709*, 2016.

[SW17] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. *SocialNLP 2017*, page 1, 2017.

[Tar16] Natalya Tarasova. Classification of hate tweets and their reasons using svm. Master's thesis, Uppsala Universitet, 2016.

[THL+16] Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. A dictionary-based approach to racism detection in dutch social media. *arXiv preprint arXiv:1608.08738*, 2016.

[Tho16a] Neil Thompson. *Anti-discriminatory practice: Equality, diversity and social justice*. Palgrave Macmillan, 2016.

[Tho16b] Annie Thorburn. Hate speech ml. Available in https://github.com/anniethorburn/Hate-Speech-ML , accessed last time in May 2017, 2016.

[TT13] Luis Torgo and Maintainer Luis Torgo. Package 'dmwr'. *Comprehensive R Archive Network*, 2013.

[Twi17] Twitter. The twitter rules. Available in https://support.twitter.com/articles/, accessed last time in February 2017, 2017.

[UCS16] UCSM. Iwg hatespeech public. Available in https://github.com/UCSM-DUE/, accessed last time in February 2017, 2016.

[VF05] Michel Verleysen and Damien Francois. The curse of dimensionality in data mining and time series prediction. In *International Work-Conference on Artificial Neural Networks*, pages 758–770. Springer, 2005.

[Was16] Zeerak Waseem. Hate speech twitter annotations. Available in https://github.com/ZeerakW/hatespeech, accessed last time in February 2017, 2016.

[Wen15] Mike Wendling. 2015: The year that angry won the internet. Available in http://www.bbc.com/news/blogs-trending-35111707, accessed last time in February 2017, 2015.

[WH12a] William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26. Association for Computational Linguistics, 2012.

[WH12b] William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26. Association for Computational Linguistics, 2012.

[WH16] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of NAACL-HLT*, pages 88–93, 2016.

[WZH01] Ke Wang, Senqiang Zhou, and Yu He. Hierarchical classification of real life documents. In *Proceedings of the 2001 SIAM International Conference on Data Mining*, pages 1–16. SIAM, 2001.

[Yah17] Yahoo! Webscope datasets. Available in https://webscope.sandbox.yahoo.com/, accessed last time in February 2017, 2017.

[Yar94] David Yarowsky. Decision lists for lexical ambiguity resolution: Application to accent restoration in spanish and french. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 88–95. Association for Computational Linguistics, 1994.

[You17] Youtube. Hate speech. Available in https://support.google.com/youtube/answer/2801939?hl=en, accessed last time in February 2017, 2017.

[YWX16] Shuhan Yuan, Xintao Wu, and Yang Xiang. A two phase deep learning model for identifying discrimination from tweets. In *International Conference on Extending Database Technology*, pages 696–697, 2016.

[Zoo12] Matthew Zook. Mapping racist tweets in response to president obama's re-election. Available in https://www.theguardian.com/news/datablog/2012/nov/09/mapping-racist-tweets-president-obama-reelection, accessed last time in May 2017, 2012.

REFERENCES

Appendix A

Graph of classes

We present in this appendix the complete graph of classes used for annotate this dataset (Figure A.1).

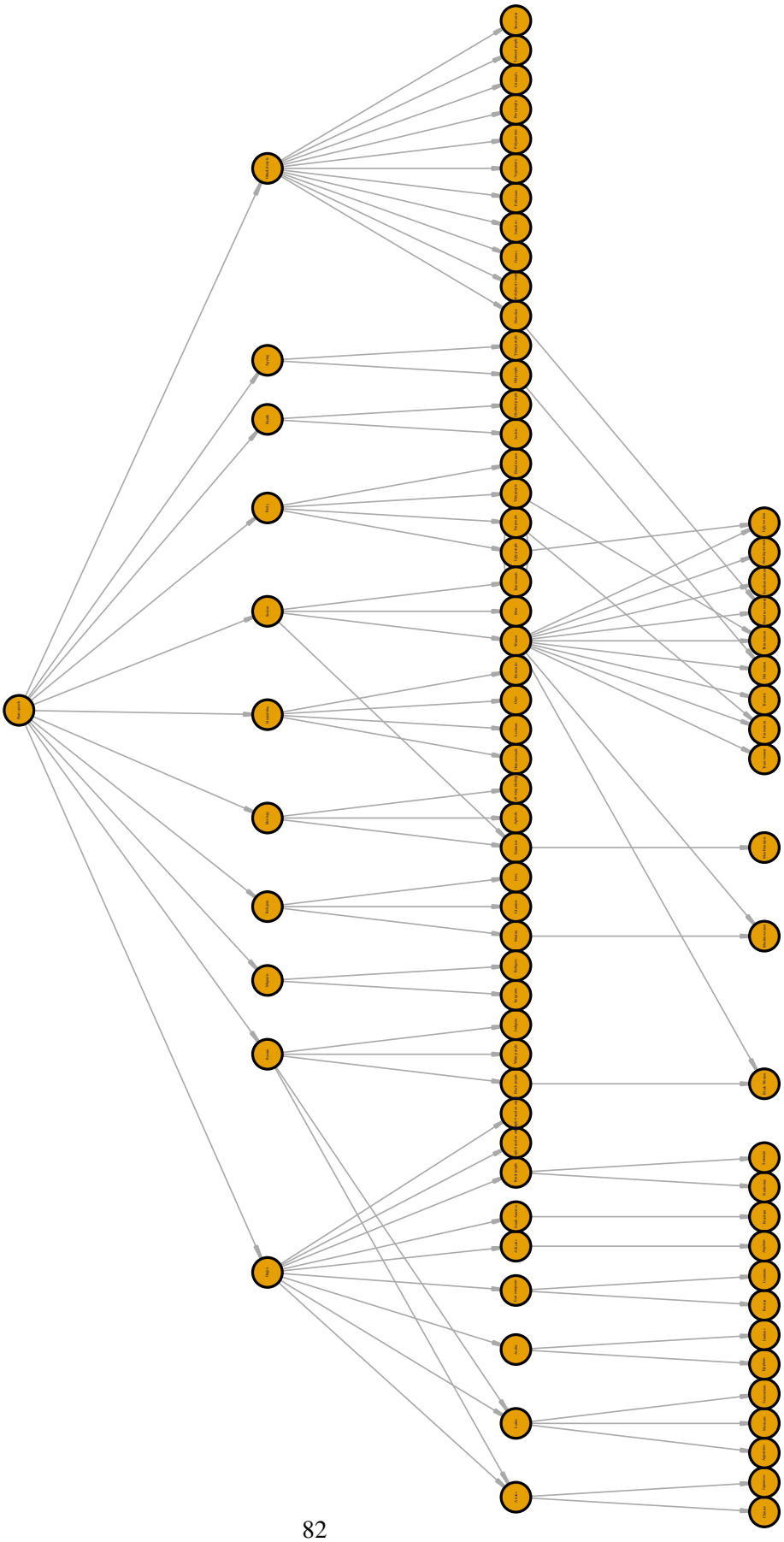

Figure A.1: Graph of classes used for annotate the dataset in Portuguese.

Appendix B

List of search instances

We present in this appendix the list of search instances that we used in our study (Figure B.1).

Table B.1: List of profiles and words used for the messages search.

Profiles	Words
OcaralhoAA4	#renovarPortugal
gentebranca1	#ESeFosseConsigo
HumorNegroo	José Pinto-Coelho
DireitaBrasil	nao queremos chineses
direitafalando	branco
Adolfo_Dias_I	nao angola
OPauloAlmeida	volta para a tua terra
BrothersDireita	volta para angola
Justiceiro_Sujo	gordo feio
agnaldoregojr	gorda feia
Bolsotrix	chines
realDTrumpBR	fufas
JOAQUINVOLTOU	sapatao
MaBolsonaria	#diadoorgulhohetero
Bolsonaopressor	#LugarDeMulherENaCozinha
editorahumanas	#MeuAmigoSecreto
otmar03	#PrimeiroAssedio
heterosim	#NaoAIdeologiaDeGenero
homemqehhomem	#MasterChefBR mulher na cozinha
homemqehomem	mulher burra
HQEH	estupro 33
homemdeverdade	#NaoAceitamosRefugiados
clubedomacho	pnr refugiados
verdadehomem	nao refugiados
MdeVerdade	refugiado terrorista
pnr	muçulmano
jpintocoelho60	Islao
NAOKAHLO	nao a burca
marisa_lobo	nao refugiados
	#foraheterofobia

List of search instances

Appendix C

Annotation instructions

The goal of this task is to mark messages as containing hate speech or not. For that task it is important to define hate speech:

- Hate speech is language that incites violence or hate against certain people or that attacks or diminishes them, 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 humour can be used.

To make it even more clear, on one hand, violence can occur physically and explicitly, however, on the other hand it can also be subtle. This is the case when stereotypes are reinforced, giving a justification to discrimination and negative bias towards these groups. Therefore, we consider that all subtle forms of discrimination, even jokes, must be marked as hate speech.

A message has hate speech when two conditions are met:

- A verbal attack occurs.
- The target of the attack is characterized as belonging to a specific group.

Specific rules and examples are presented in the table C.1.

The presented rules point out that we aim to have a more inclusive and general definition about hate speech than some other perspectives found in literature. This is the case because we want to be able to better describe subtle forms of discrimination amongst the internet and on social networks.

Finally, each message should be marked with the targeted group name or none. More than one group can be used also. Some more examples are also provided (Table C.2).

Particular cases:

- Sometimes it is not possible to explicitly identify a group in the messages: “I will always have hetero pride.” In this cases the name of the prejudice should be used: “homophobia”.

Table C.1: Rules for annotation with examples and classification.

Rule	**Example Message**	**Class**
Usage of disparaging terms and racial epithets with the intent to harm must be considered hate speech.	Dykes are crazy.	Lesbians
However, in a discussion of the words themselves such expressions might be acceptable.	Dyke is an offensive word for a lesbian.	None
We can say that if a text "uses a sexist or racial slur" it has hate speech.	Thanks, fat ugly woman.	Fat, ugly, woman
Sometimes these words are used by a speaker who belongs to the targeted group, in order to show pride for belonging to the group. For our purpose, and if there is no contextual clue about it, such terms are categorized as hateful.	I always have been a dyke.	Lesbians
Also, references to an organization associated with hate crimes does not by itself constitute hate speech. For instance the name "Ku Klux Klan" is not hateful, as it may appear in historical articles or other legitimate communication.	Ku Klux Klan is a group in the United States.	None
However, while the endorsement of organizations that promote hate speech does not constitute a verbal attack on another group, in the scope of this work we define that this must be marked as hate speech.	I'm not macho, but sexist jokes are very funny.	Sexism
References to behaviors that contain hate speech should be marked as hate speech for descriptive purposes.	#mysecretfriend likes to see lesbians kissing but if it is gay there is sin and god does not like.	Homossexuals
Besides, calling attention to the fact that an individual belongs to a group and invoking a well known and disparaging stereotype about that group is hate speech as well.	Woman's place is in the kitchen!	Woman
Making generalized negative statements about minority groups as in "the refugees will live off our money" is hate speech, due to the incitation of a negative bias towards the group.	Refugees are more like "Rapefugees"	Refugees
However it is also important to point out that the use of some words like "black", "white", "filthy", or other, is marked as hate speech only in some circumstances. Outside of context, these words bear no racial undertones of their own.	I love black and white.	None
Members of religious groups are protected, religion itself is not.	1) Islamism is submission. 2) Mosque in Lisbon? No!	1) None 2) Islamites
Speaking badly about countries (e.g France or Germany) is allowed in general, however condemning people on the basis of their nationality is not.	Brazil deserves it.	None
Finally, hate speech can also occur when the statement about the superiority of the in-group are made.	Employment: priority to the Portuguese!	Immigrants
Shows support of problematic hashtags.	Every day I have #Heteropride.	Homophobia

- Sometimes more than one group is targeted in the message. More than one label should be used.

Table C.2: More examples of hate speech classification

Message	**Class**
Suck it, you filthy old man!	Old people
Dwarf = Penguin	Dwarf
Left people, human garbage.	People with left wing ideology
Feminicide is victimhood.	Woman
Boys don't cry!	Man
Nobody borns gay.	Gays
To be a woman is to pee sitting.	Trans woman
I saw things I wanted to unsee.	None

Appendix D

N-grams results in Portuguese

We present in this appendix the original Table D.1 with the results in Portuguese.

Table D.1: Top-10 n-grams more common in the classes “Hate speech”, “None”, “Health”, “Homophobia”, “Ideology”, “Origin”, “Racism”, “Religion”, “Sexism” and “Other-lifestyle”, in Portuguese.

none	**hate speech**	**health**	**homophobia**	**ideology**
branco	mulher	musica	sapatão	feminista
pnr	gorda	ouvir	fufas	feministas
angola	burra	donos	sapatao	esquerdopatas
sobre	feia	gostava	q	feminazi
bolsonaro	é	doença	orgulho	ser
esquerda	sapatão	qm	viado	feminismo
trump	ser	burro	hetero	pra
todos	fufas	muitos	mundo	gorda
grande	q	caralho	gay	é
agora	homem	galera	gays	pai
origin	**racism**	**religion**	**sexism**	**other-lifestyle**
angola	racismo	islão	mulher	bandido
latino	branco	muçulmanos	gorda	é
nordeste	negro	muçulmano	burra	bom
entrar	negros	X5	feia	morto
povo	é	europa	é	criminosos
assim	pode	aqui	ser	rir
pode	preto	segura	homem	pode
luz	pq	bomba	q	gente
terrorista	existe	torna	dia	pra
posso	nada	mesquita	pra	namorar