DETECCIÓN DE *HATE SPEECH* ONLINE UTILIZANDO MACHINE LEARNING

ONLINE HATE SPEECH DETECTION USING MACHINE LEARNING

TRABAJO FIN DE GRADO
CURSO 2021-2022

AUTORA
ELA KATHERINE SHEPHERD ARÉVALO

DIRECTORES
GONZALO MÉNDEZ POZO
PABLO GERVÁS GÓMEZ-NAVARRO

GRADO EN INGENIERÍA INFORMÁTICA
FACULTAD DE INFORMÁTICA
UNIVERSIDAD COMPLUTENSE DE MADRID

DETECCIÓN DE *HATE SPEECH* ONLINE UTILIZANDO MACHINE LEARNING

ONLINE HATE SPEECH DETECTION USING MACHINE LEARNING

TRABAJO DE FIN DE GRADO EN INGENIERÍA INFORMÁTICA
DEPARTAMENTO DE INGENIERÍA DEL SOFTWARE E INTELIGENCIA ARTIFICIAL

AUTORA
ELA KATHERINE SHEPHERD ARÉVALO

DIRECTORES
GONZALO MÉNDEZ POZO
PABLO GERVÁS GÓMEZ-NAVARRO

CONVOCATORIA: SEPTIEMBRE - 2022
CALIFICACIÓN:

GRADO EN INGENIERÍA INFORMÁTICA
FACULTAD DE INFORMÁTICA
UNIVERSIDAD COMPLUTENSE DE MADRID

DÍA DE MES DE AÑO

INSCRIPTION

To myself, for pushing through all these years without even knowing where I was going.

ACKNOWLEDGEMENTS

RESUMEN

Un resumen en castellano de media página, incluyendo el título en castellano. A continuación, se escribirá una lista de no más de 10 palabras clave en inglés

Palabras clave

ABSTRACT

An abstract in English, no more than a half page, including the title in English. Below, a list with no more than 10 keywords.

Keywords

Artificial Intelligence, Machine Learning, Hate Speech, NLP, Social media, Twitter

CONTENT INDEX

FIGURE INDEX

TABLE INDEX

Chapter 1 - Introduction

In this first chapter I will explain the motivation behind this project and the main objectives I intend to reach at the end of it. I will also describe how this document is divided, briefly explaining each section.

1.1 Motivation

Hate speech is describes as public speech that expresses hate or encourages violence towards a person or group based on a social characteristic (gender, race, sexual orientation, etc.) It is a pandemic that has been rampant around the world for a very long time, and in recent decades the rising od social media has made it easier to bully, mock and degrade oppressed social groups Having the freedom to express every thought and feeling on the Internet is a double edged sword; when everybody is able to communicate everything with everybody with no limits, hateful communities are formed, new insults arise (Pascoe & Diefendorf, 2019) and targeted attacks can be planned (Bliuc, Faulkner, Jakubowicz, & McGarty, 2018). It's even been studied that online hate speech can predict violence: (Blake, O'Dean, Lian, & Denson, 2021)'s research shows a correlation between misogynistic tweets and domestic violence.

As a bisexual person on the Internet who has presented as a woman all their life, I'm no stranger to hate speech spoken by peers or people online, either towards me or towards my loved ones. I believe that it is important to overview the content that is posted online, not for the purpose of censoring, but to avoid young people experiencing trauma from a young age online and to dodge the possibility of other people falling into the same mentality.

In order to achieve this, the first step is to learn how to properly detect hate speech online. This means both aggressive and non-aggressive hate speech. During this project I will use different machine learning (ML) approaches to classify hate speech in different languages: English, Spanish, Italian, Portuguese, and finally I will also use the same ML algorithms to create a multilingual classifier in the three mentioned Mediterranean languages. Afterwards I will compare the results to see what algorithms can detect hate speech better, and how language and data volume can play a role in the accuracy.

1.2 Objectives

My main objectives that I wish to accomplish in this project are:

- Gather data to create a well-sized, classified corpus of internet posts that may or may not present hate speech
- Use a different set of ML approaches to classify the data
- Compare the different outcomes and study which models/approaches could be further used in future work.

1.3 Document structure

[Pending: Will write when rest of the chapters and appendixes are done]

Chapter 2 - State of the art

In this chapter I will elaborate on the most important sides of the current technological and academical state of the project's topic.

2.1 Machine Learning

2.1.1 Logistic Regression

2.1.2 Naive Bayes

2.1.3 Support Vector Machines

2.2 Deep Learning

2.3 Hate speech detection

Automated hate speech detection is a NLP that has been gaining popularity in recent years. A 2019 survey on Automatic misogyny detection (Shushkevich & Cardiff, 2019) explains two main approaches for it:

- Using classical Machine Learning algorithms such as Support Vector Machines, Naïve Bayes and Logistic Regression
- Using neural networks. Some examples cited in the article is (Zhang & Luo, 2018) and (Park & Fung, 2017) and (Badjatiya, Gupta, Gupta, & Varma, 2017)'s use of CNNs and (Goenaga, y otros, 2018) and (Badjatiya, Gupta, Gupta, & Varma, 2017)'s approaches with RNNs (specifically with LSTMs).

Of course, this project's approach is similar to this last one. Although none of the examples I just named are on Spanish data or BERT, this doesn't mean there's no research done about Misogyny detection using such language and model. For example, (Plaza-Del-Arco, Molina-González, Ureña-López, & Martín-Valdivia, 2020)'s paper uses Spanish data from the AMI shared task. However, none of their approaches include BERT, but they do explain that they wish to explore models like ELMO and BERT as future work. The

case is the same in other studies (García-Díaz, Cánovas-García, Colomo-Palacios, & Valencia-García, 2021).

It's important to note that shared tasks have helped a great deal with sexism detection. The most cited one in my research was IberEval's 2018 shared task con Automatic Misogyny Identification (AMI) (Fersini, Rosso, & Anzovino, 2018) . Other notable shared tasks in Spanish on this same topic is SemEval's 2019 task on Multilingual detection of hate speech against immigrants and women on Twitter (Basile, y otros, 2019) and IberFEL's 2021 EXIST shared task (Rodríguez-Sánchez, y otros, 2021).

This lastly mentioned task, the EXIST task, has some interesting studies done on it where the main approach is with BERT (Butt, Ashraf, Sidorov, & Gelbukh, 2021) (Paula, Silva, & Schlicht, 2021). However, they aren't the only projects on hate speech detection using BERT. Even though their goal was and not sexism detection, (Caselli, Basile, Mitrović, & Granitzer, 2020) re-trained BERT in order to be able to detect abusive language in short, English text, using comments from an important internet forum, Reddit.

Chapter 3 - Corpus

In order to classify using the array of models I have chosen; I searched for datasets on hate speech. This could be general hate speech or a specific kind of hate speech (racism, homophobia, sexism, etc).

For each language in the corpora, I didn't take into consideration the geolocation of the posts. This means that, for example, European Spanish and South American Spanish, Brazilian Portuguese and European Portuguese, and British English and American English could be analysed together.

Dataset id	**Languages**	**N° of total rows**	**Percentage of text with hate speech (*before preprocessing)***
EXIST	English, Spanish	English: 5644 Spanish: 5701	English: 49.5% Spanish: 50.24%
AMI 2020	Italian	5409	47.09%
HatEval	English, Spanish	English: 13000 Spanish: 6600	English: 42.08% Spanish: 41.5%
Automated Hate Speech Detection	English	24783	5.77%
IHSC	Italian	6928	18.63%
Hierarchically-Labeled Portuguese Hate Speech	Portuguese	5668	19.85%
Hateful Symbols or Hateful People?	English	16909	31.63%
Are You a Racist or Am I Seeing Things?	English	6909	18.18%
HaSpeeDe 2	Italian	6837	40.46%
ToLD-BR	Portuguese	21000	1.79%
OffComBR	Portuguese	1033	19.55%

Hate speech dataset from a white supremacist forum	English	10944	10.93%
The Gab Hate Corpus	English	27546	8.52%

Table 1. List of datasets used in this project

In this chapter I will go through each dataset used in this project; describing its structure and the processing done to them with the objective of unifying and cleaning it for the experiments, in such as way that all datasets are reduced to two columns: the text of the post and whether it contains hate speech (1) or not (0).

It's important to note that for each dataset with train and test sets, both sets (and, if were present, validation set) were joined after processing in order to have a mixed corpus of all corpora, which would later be split for the classification.

3.1 EXIST dataset

IberLEF is a shared evaluation campaign for NLP systems in Iberian languages, such as Spanish and Portuguese. Their goals is to encourage research in this field so more state-of-the-art tasks are done in these languages. Every year there is a call for different task proposals, and those who are interested can apply, making it an international collaboration ending with interesting results and interested participants.

Last year, in 2021, one of the shared tasks was EXIST: sEXism Identification in Social neTworks (Rodríguez-Sánchez, y otros, 2021). Their objective was to be able to detect sexism in social media posts; from small micro-aggressions to violent misogyny. The data that would be classified was a list of tweets, from the social network Twitter; and gab posts, from the far-right social network Gab. All this text was obtained by collecting many sexist terms used on the internet, and subsequently extracting tweets and gab posts that used those expressions.

Participants were asked to classify the data in accordance with two tasks:

- **Sexism identification**: A binary classification of whether a text was sexist or not. The "degree" of sexism is not important.
- **Sexism categorization**: If a text is sexist, the goal of this task was to determine what kind of sexism was present. Some of these labels were:

Since this project uses more than one dataset to form the corpus, only the EXIST sets were classified by type of sexism, so I didn't make use of the second task, and focused on the first, binary-classifying task.

test_case	id	source	language	text	task1	task2
EXIST2021	10280	gab	es	puta madre	non-sexist	non-sexist
EXIST2021	10534	twitter	es	No puedo más con las zorras	sexist	misogyny-non-sexual-violence
EXIST2021	007019	twitter	en	At what point did I slut-shame anyone? I said that wasn't how I got into uni.	non-sexist	non-sexist

Table 2. Sample of original EXIST dataset

3.2 AMI dataset

Evalita is an annual evaluation campaign on NLP in Italian. It's been organizing shared tasks since 2007 and is endorsed by the Italian Association for Artificial Intelligence and the Italian Association for Speech Sciences. (Fersini, Nozza, & Rosso, 2020) presented a shared task on Automatic Misogyny Identification, shortened as AMI. However, this wasn't the first time it has appeared in a shared task. Both IberEval and Evalita had an AMI shared task in 2018. These tasks have data in English, Spanish and Italian. Unfortunately, I could only obtain the train and test set from the 2020 shared task, which is in Italian.

The goal of this task was not only to identify misogyny, but to recognize whether or not a piece of text is aggressive. Therefore, the data, apart from the text, had two columns:

- **Misogynous**: A binary classification of whether a text was sexist or not. The "degree" of sexism is not important.
- **Aggressiveness**: A binary classification of whether a text was aggressive or not.

I only kept the information from the first mentioned column.

3.3 HatEval dataset

SemEval is a collection of research workshops on NLP whose aim, much like IberLEF's, is to advance the state of the art in this field and to create datasets for many shared tasks on natural language semantics. SemEval's tasks are an annual event, and in 2019, 13 new tasks on semantic evaluation were announced, the fifth being HatEval: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. (Basile, y otros, 2019)

This task's goal was to detect hate speech directed at two vulnerable social groups: immigrants and women. The data to classify in this task was in English and Spanish.

Here, participants were asked to identify several aspects of the tweets in two different tasks:

- **Hate Speech Detection against Immigrants and Women**: Whether or not a tweet contained hate speech towards women or immigrants.
- **Aggressive behavior and Target Classification**: Describing whether the hate speech in the text is aggressive, and also identifying if the hate speech is directed to a specific person or to a group of individuals.

To unify the datasets, I only used the values in the column describing the values for the first task.

3.4 Automated Hate Speech Detection dataset

(Davidson, Warmsley, Macy, & Weber, 2017) created a dataset that categorized thousands of tweets into three categories: hate speech, offensive text and neither. It's important to make this distinction, because many times a slur can be used in a way that isn't hate speech. Black people use racial slurs used against them as slang between each other, and sometimes this can be offensive, but not hate speech.

Davidson's team had workers from CrowdFlower manually annotate the tweets. In the end, the dataset available to the public was formed by these columns:

- **Count:** Number of CrowdFlower users who annotated the tweet

- **Hate_speech:** Number of CrowdFlower users who considered the tweet to be hate speech
- **Offensive_language:** Number of CrowdFlower users who judged the tweet as offensive
- **Neither:** Number of CrowdFlower users who described the tweet as neither offensive nor non-offfensive
- **Class:** Class label for majority of CrowdFlower users (0: hate speech; 1: offensive text; 2: neither)
- **Tweet:** Full text of the tweet

Since the "class" column, which is the only one I wanted to keep for the experiments, wasn't a binary value, I changed the values of the columns as so:

- $0 \rightarrow 1$
- $1 \rightarrow 0$
- $2 \rightarrow 0$

3.5 IHSC dataset

(Sanguinetti, Poletto, Bosco, Patti, & Stranisci, 2018) made the contribution of an Italian dataset about hate speech in tweets about, mainly, immigrants. The dataset has 6 different columns:

- **Tweet_id:** The Twitter ID of the tweet
- **Hs:** Indication of whether a tweet contained hate speech or not (yes/no)
- **Aggressiveness:** Value describing if a tweet has the intention of being aggressive or harmful (no/weak/strong)
- **Offensiveness:** This column indicated if a tweet was hurtful (no/weak/strong)
- **Irony:** Indication if a tweet contained sarcasm, satire or irony to imply a certain message (yes/no)

- **Stereotype:** Whether a tweet implicitly or explicitly falls into beliefs society has about different groups of people (yes/no)

Since the text of the tweets aren't available in this dataset, it was necessary to use Twitter's API, Tweepy, to extract it. All rows where a tweet couldn't be extracted (the account of the user was suspended, the tweet was deleted, etc.) were eliminated.

I only made use of the column that indicated if a tweet contained hate speech or not.

3.6 Hierarchically-Labeled Portuguese Hate Speech dataset

(Fortuna, Silva, Soler-Company, Wanner, & Nunes, 2019) built a Portuguese dataset for hate speech detection research. They annotated their obtained tweets in two ways: binary and hierarchical. In the binary annotation annotators had to classify as hate speech (1) o not hate speech (0). On the other hand, the hierarchically annotated dataset had a large number of columns; each one of them being a group of people a tweet could potentially be targeted at. These classes work as a tree-like structure, where one class can have one or more child classes ("Asians" is a child class of "Racism").

Even though I could've simply used the binary annotation, I wanted to do a more complex analysis of the different classes present in the hierarchal annotation, because I will not count everything categorized here as hate speech as so.

Here are the classes I didn't count as hate speech, and so, would be deleted from the dataset (I also deleted the global existing parent class "hate speech", since it wasn't necessary):

- Body
- Ideology
- Agnostic
- Criminals
- Journalists
- Left wing ideology
- Men Feminists
- Old people

- Polyamorous
- Russians
- Street artist
- Ukrainians
- Vegetarians
- White people
- Young people
- Men
- East Europeans
- Thin people
- Ageing

I did so following my definition for hate speech in this project, which I talk about a little more on appendix B. In the cases where I was on the fence (polyamorous, old people, east Europeans), I've deleted these columns because text targeted towards them would be an extremely small percentage of the data.

But knowing how many of these classes have parent classes, that would mean that if we just deleted these columns, their remaining parent classes would still have a positive value indicating hate speech stemming from their deleted child. So, before deleting these columns, I set their parents' (plus their own) column values as 0.

I then unified all these columns as a new, single "hate speech" column, that would have a 1 if at least one of the remaining columns had the value of 1, and a 0 in the other case.

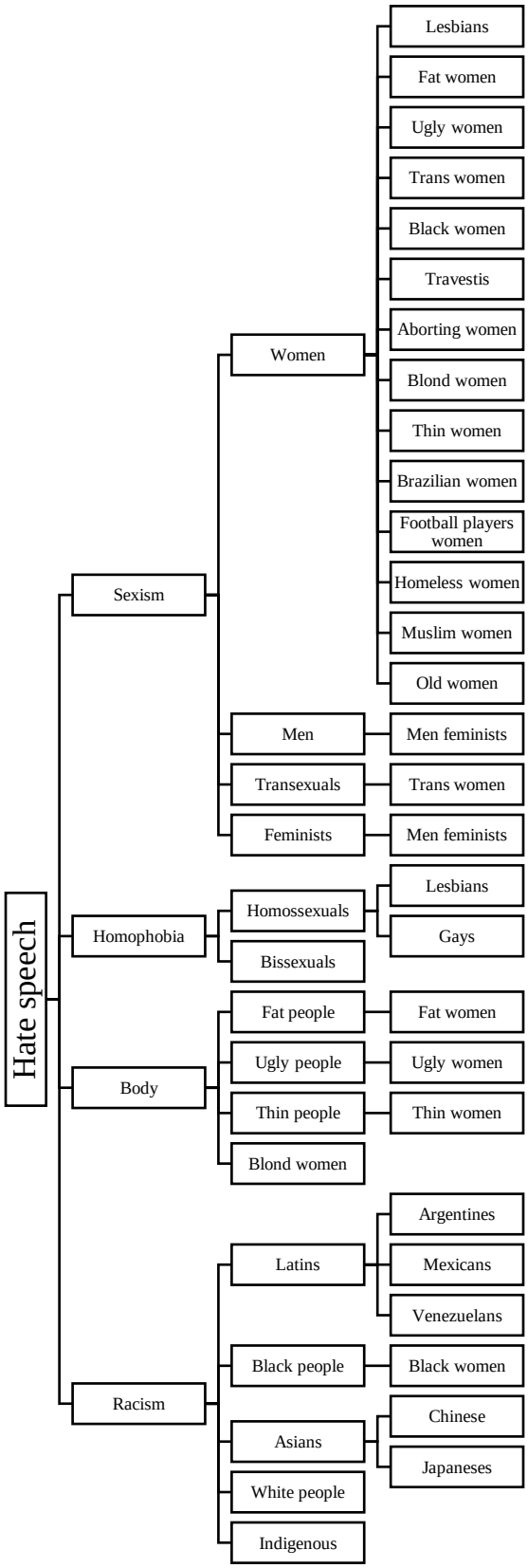


Figure 1. Data structure for Hierarchically-Labeled Portuguese Hate Speech set, part I

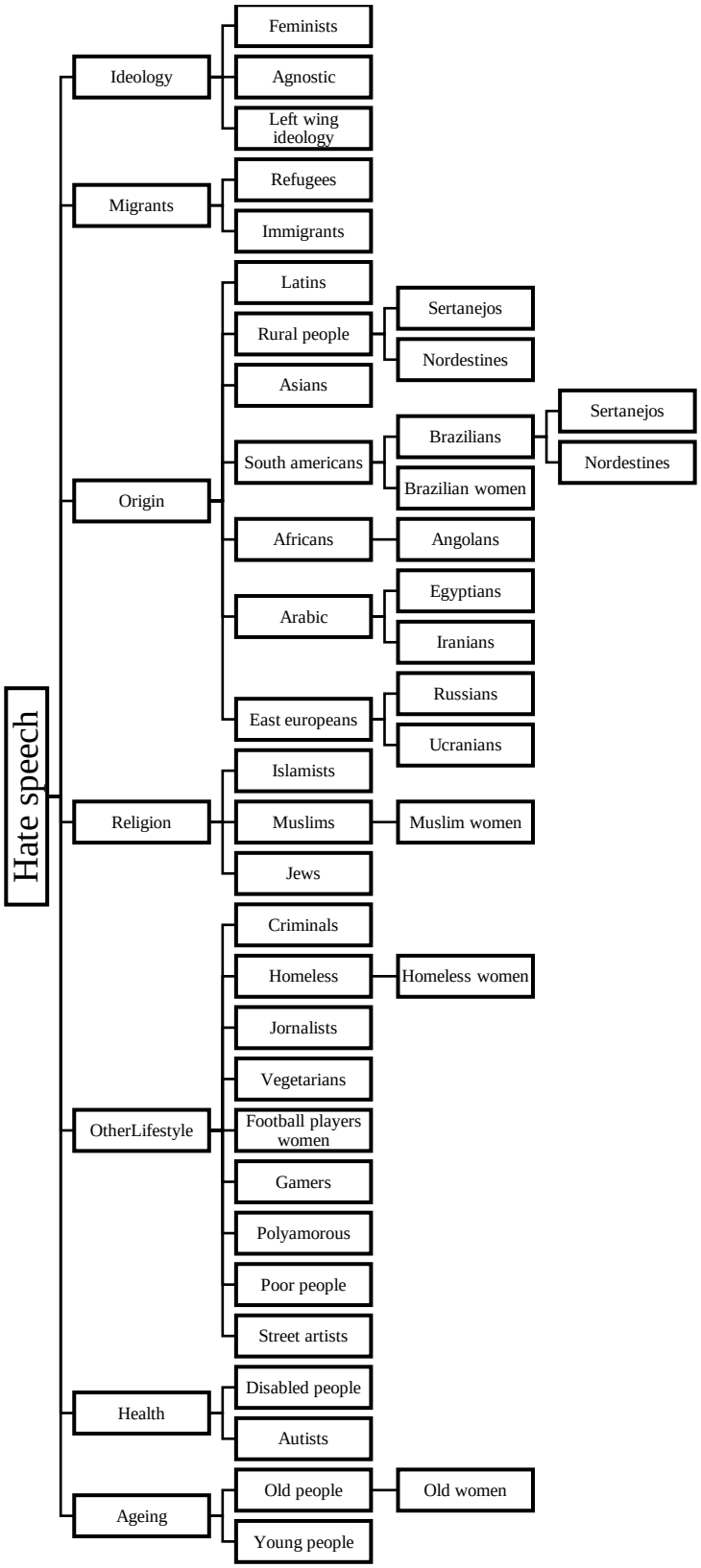


Figure 2. Data structure for Hierarchically-Labeled Portuguese Hate Speech set, part II

3.7 Hateful Symbols or Hateful People? dataset

(Waseem & Hovy, Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter, 2016) created a dataset that detected hate speech which contained two columns: the first being a tweet ID and the second being a string that indicated the hate speech that the tweet contained ("sexism", "racism", "none"). I used Tweepy just like in the IHSC dataset to return the tweet text. I deleted all rows where the text couldn't be extracted, and finally, I manually checked the classification of all tweets that overlapped with the dataset I will talk about next.

3.8 Are You a Racist or Am I Seeing Things? Dataset

(Waseem, Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter, 2016) wanted to look more into the annotation of hate speech, and how different ways of annotating data can influence how good the annotations are. The tweets were classified by amateur annotators from CrowdFlower and by expert annotators – anti-racist and feminist activists that has much knowledge on the subject. In his conclusions we see that having intimate knowledge about a subject helps a great deal when it comes to detecting hate speech related to that topic.

The data that Waseem used some of the tweets from (Waseem & Hovy, Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter, 2016), plus many new ones. Therefore, I deleted all of the overlapping tweets, and double checked the classification on the mentioned dataset.

For each column, the set contained many columns: the first being the Twitter ID of the tweet. Secondly, there is a "Expert" column, indicating the classification made by the expert annotator assigned to the text. Finally, we find a varying number of "amateur" columns, which indicate the classification of the tweet by one of the amateur annotators that judged the tweet. The tweets could be classified as four different categories: the ones used in Waseem and Hovy's paper plus "both", indicating that a tweet contained racism and sexism.

In order to create a single binary classification column, I considered the tag "neither" to be a 0, and in the rest of the cases I would tag the tweet as 1. On the other hand, to unify all different columns that annotated a tweet, I considered the tweet to contain hate speech if there were more 1s tan 0s, and vice versa (e.g.: If the expert annotator plus two amateur annotators classified the tweet as 1, and only one amateur annotator classified it as 0, my column would contain a 1). Finally, I wanted to check the tweets were the majority wasn't the same as the Expert classification, so I manually classified those myself.

3.9 HaSpeeDe 2 dataset

(Sanguinetti, y otros, 2020)

3.10 ToLD-Br dataset

(Leite, Silva, Bontcheva, & Scarton, 2020) proposed a large-scale dataset for detecting toxicity in tweets in Brazilian Portuguese. The dataset was created with the help of 129 annotators for fine-tuning monolingual and multilingual BERT models.

In this set, "toxicity" isn't equivalent to hate speech. For every tweet, there are 6 column each indicating a different aspect of toxicity: Homophobia, Obscene, Insult, Racism, Misogyny, and Xenophobia. Insulting and obscene tweets are definitely toxic, but don't count as hate speech, since the target isn't so because of a social category. The categories can overlap, but in the cases where they don't, I've decided to classify those tweets that are exclusively obscene and/or insulting them as not hate speech.

Every column contains values between 0 and 3, signifying the number of annotators that considered that the tweet belonged to such category (each tweet was labelled by 3 annotators). Since I'm working with binary values, if a row has the value 1 in a column, that value would be replaced by a 0. Only those tweets with where the majority of annotators have judged them to be hateful would be categorized as hate speech.

3.11 OffComBR

Since hate speech detection in Portuguese (and specifically, Brazilian Portuguese) has little research, (Pelle & Moreira, 2017)'s contributions where helpful, by, among other results, creating a Portuguese dataset with hateful and non-hateful comments from the Brazilian news site g1.globo.com, specifically from the politics and sports sections of the site, since pieces of news from such sections contained the majority of hateful comments.

For each comment, the data would contain an id, a class or classification, and the comment's full text. There could only be two kinds of category a comment could be in, hateful or non-hateful, written as "yes" if the comment was hateful or "no" if otherwise.

Two datasets where developed. The first contained all of the extracted comments, and the assigned class was the one picked by at least two out of the three judges to annotate the comment. The second dataset, which I used for this project, only contained the comments where all annotators agreed on the class. This way, there might be less data, but more accuracy on it. All I needed to do was to change the classes names from "yes" and "no" to 1 and 0.

3.12 Hate speech dataset from a white supremacist forum dataset

(Gibert, Perez, García-Pablos, & Cuadros, 2018) created a hate speech dataset based on posts on the white supremacist, neo-Nazi Internet forum, Stormfront. In this dataset, posts were classified into four different categories:

- **Hate**: Text that contains hate speech
- **No hate**: Text that doesn't contain hate speech
- **Relation**: Text that, by itself, doesn't convey hate, but combining it with other sentences in this category, does
- **Skip**: Sentences in other languages or so neutral that it doesn't enter in any of the other categories

To only have two values, I deleted all the "skip" columns; and as for the "relation" posts, I wanted to manually classify them myself, because I had the suspicion that some of them could potentially contain hate speech by itself, in a more subtle way.

And this was the case. On some of the posts categorized as "relation", the users affirm false, hateful stereotypes or refer to a group of people they despise in a pejorative way, even through they're not insults or slurs.

For example, the post "The same way Jews run the government ." is a sentence that defends an antisemitic stereotype of Jews being powerful overlords.

In another case, in the phrase "Maaaaany pinders and Asians here .....", the user uses "pinder" as an ethnic slur against East Asians. But use of slurs alone isn't enough to categorize text as hateful. As (Bianchi, 2014) explains, many marginalized communities have reclaimed slurs directed at them and use them in a friendly context between them. The most known example is the use of nigg*r within the black community. However, it's important to know what slurs have been reclaimed, because if it isn't the case, it would certainly mean that the person is simply using it in a negative way (which is the case in the example, obviously, since this is taken out of a neo-Nazi forum).

A final example I want to show is the post "(Includes : one pair of baggy pants , one pistol , a set of golden grills , a looting guide titled ' But I dendu nuffin ' , and one race card)". There's no explicit reference to a group to stereotype, nor are there any slurs present; but we do find the term "dindu nuffin", which is an anti-black expression that originated in 2014. This is quite recent, and I point this example out to explain that consciousness about rapidly changing hateful speech on the Internet is essential in these tasks.

Taking all of this into consideration, I manually classified the "relation" posts into hate speech or not hate speech.

3.13 The Gab Hate Corpus dataset

(Kennedy, y otros, 2022)'s data is a set formed by posts from Gab, and the typology that they use to categorize the posts in different column goes as so:

- **HD**: Meaning "assault on human dignity"
- **CV**: Meaning "call for violence"
- **VO**: Meaning "Vulgarity/Offensive language directed at an individual"

I discarded the columns "CV" and "VO", since text that endorses violence and text that is vulgar aren't necessarily hate speech. For example, one could call for violence against a politician for purely political reasons, and not because of their race, gender, religion, body, etc.

The "HD" column is all the information I needed from this dataset, since this column checks if the text contains superiority over a certain social group by using slurs, stereotypes or references.

3.14 General ML preprocessing

After preparing each dataset individually, I studied and applied some NLP pre-processing techniques used in (Arriba, Oriol, & Franch, 2021) to clean up the text.

URLS and mentions and Retweet indicators

URLs and mentions to other users (words beginning with @) were removed in order to not incorrectly teach the model that mentioning a certain user could determine if the text was hateful or not.

Also, if a tweet began with the letters "RT" (meaning that the tweet was a Retweet), they would be deleted.

Lemmatization

Stopwords

Hashtags and emojis

For emojis and hashtags, I decided to create three different datasets for each language, depending on how I dealt with them. The first type of datasets would have all emojis and hashtags removed. The second type would maintain them all, leaving the hashtags as they are and replacing each emoji for their "demojized" version using the emoji python library. So, for example, the 👍 emoji would be replaced by ":thumbs_up:". I didn't eliminate the colons since I saw them as good indicators for emojis.

Finally, the third group would contain only a number of emojis and hashtags, filtered out by me. In this section I will explain how I decided on the hashtags and emojis to eliminate:

Hashtags

Unlike with URLs and mentions, hashtags may show the sentiment of a tweet. For example, in the context of feminist and anti-feminist online content, if a tweet contains any of these hashtags, it's very possible the tweet is defending women's rights:

- #yositecreo: ("I do believe you", a Spanish expression used to support female victims of assault that weren't believed by peers or the justice system)
- #metoo: (An online movement where women publicized their experiences of sexual abuse)
- #niunamenos ("Not one less", a Spanish expression that demands that no more women be killed by men)

On the other hand, there are also some hashtags that are usually attributed to sexism online:

- #notallmen (An expression originated among Men's Rights Activists responding to feminists movements, commonly used to dismiss feminist talking points)
- #redpill (A term deriving from The Wachowski sisters' *The Matrix* and coined by incels describing the process in which men "realize" that they do not hold systemic power, rather, that women are the true social, economic and sexual oppressors)
- #feminazi (A pejorative term for feminists popularized by conservatives)

The case is the same with hate targeting other social groups. Of course, there may be exceptions, but I believed it was important to not dismiss such hashtags.

Because of this, I filtered all the different hashtags present throughout the corpus and checked the number of appearances of each one to *a*) work out if the hashtag could indicate whether a tweet was sexist or not b) see if there were enough appearances of it to affect the training process.

[continue to explain which specific ones were kept]

Emojis

Emojis, much like hashtags, can also influence the sentiment of a tweet, but unlike hashtags (thankfully) they can't indicate bigoted tones. But some of them can indicate certain moods:

- 😭 😟: Sad emojis can express negative feelings, but could also be used sarcastically
- 🤣 😂: Laughter can indicate positivity and are also used sarcastically
- 👏: Applause shows support or agreement

I followed the same process as hashtags: I extracted all emojis used in the data and the number of appearances to study which ones were important enough to keep in the text.

3.15 Discarded datasets

3.15.1 Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior dataset

(Founta, y otros, 2018)

3.15.2 HASOC

Chapter 4 - Experiments

4.1 Language

4.1.1 Balancing data

4.2 Model

4.2.1 Naive Bayes

Mention that we dont use gaussian because gaussian takes dense data and we use sparce data! And usually multinomial is better I believe

4.2.2 Support Vector Machines

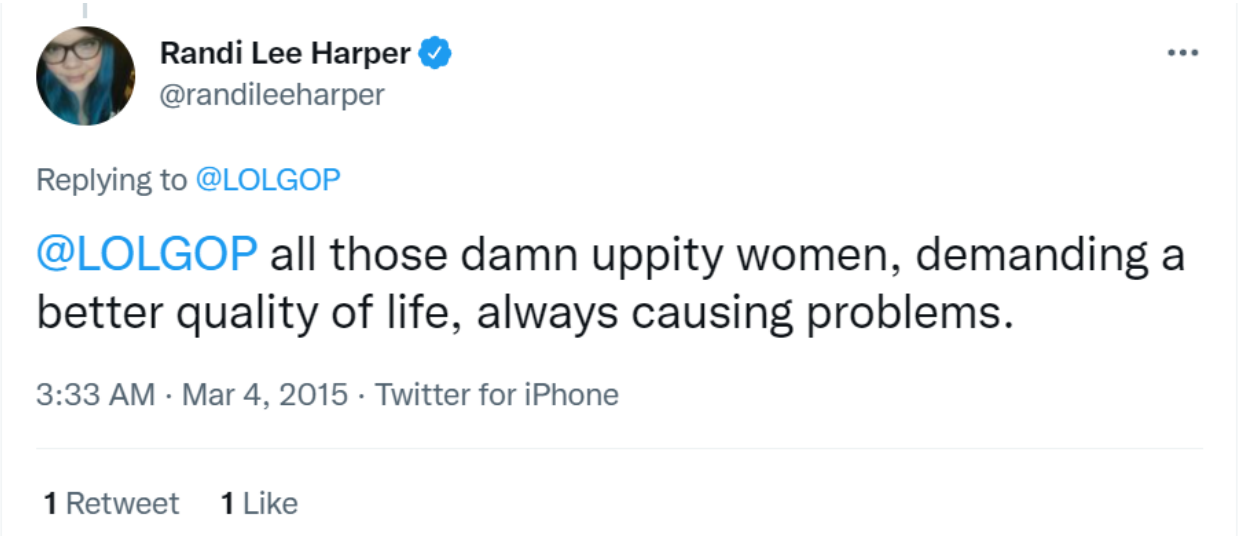
4.2.3 Logistic Regression

4.3 Train/test split

4.4 Emoji & hashtag preprocessing

4.5 Final findings

Chapter 5 - Conclusions and future work



This is sarcastic, major progress is needed to detect stuff like this, see context of full account, etc.... (edit and censor photo)

Chapter 6 - Conclusiones y trabajo futuro

BIBLIOGRAPHY

Ahluwalia, R., Shcherbinina, E., Callow, E., Nascimento, A., & Cock, M. (2018). Detecting Misogynous Tweets. En *IberEval@SEPLN*.

Arriba, A. d., Oriol, M., & Franch, X. (2021). Applying Sentiment Analysis on Spanish Tweets Using BETO.

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep Learning for Hate Speech Detection in Tweets. Obtenido de http://arxiv.org/abs/1706.00188

Bashar, M. A., Nayak, R., Suzor, N., & Weir, B. (2019). Misogynistic Tweet Detection: Modelling {CNN} with Small Datasets. En *Communications in Computer and Information Science* (págs. 3-16). Springer Singapore. doi:10.1007/978-981-13-6661-1_1

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., . . . Sanguinetti, M. (6 de 2019). SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. *Proceedings of the 13th International Workshop on Semantic Evaluation*, 54-63. Minneapolis, Minnesota, USA: Association for Computational Linguistics. doi:10.18653/v1/S19-2007

Blake, K. R., O'Dean, S. M., Lian, J., & Denson, T. F. (3 de 2021). Misogynistic Tweets Correlate With Violence Against Women. *Psychological Science, 32*, 315-325. doi:10.1177/0956797620968529

Blanco Toledano, R. (2021). Identificación de lenguaje misógino a partir de minería de textos en redes sociales. (U. N. Artificial, Ed.)

Bliuc, A.-M., Faulkner, N., Jakubowicz, A., & McGarty, C. (2018). Online networks of racial hate: A systematic review of 10 years of research on cyber-racism. *87*, 75-86. Obtenido de https://doi.org/10.1016/j.chb.2018.05.026.

Blodgett, S. L., Barocas, S., III, H. D., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. Obtenido de https://arxiv.org/abs/2005.14050

Butt, S., Ashraf, N., Sidorov, G., & Gelbukh, A. (2021). Sexism Identification using BERT and Data Augmentation – EXIST2021.

Caselli, T., Basile, V., Mitrović, J., & Granitzer, M. (2020). HateBERT: Retraining BERT for Abusive Language Detection in English. Obtenido de https://arxiv.org/abs/2010.12472

Chen, X., Wu, Y., Wang, Z., Liu, S., & Li, J. (2020). Developing Real-time Streaming Transformer Transducer for Speech Recognition on Large-scale Dataset. Obtenido de https://arxiv.org/abs/2010.11395

Davidson, T., Bhattacharya, D., & Weber, I. (8 de 2019). Racial Bias in Hate Speech and Abusive Language Detection Datasets. *Proceedings of the Third Workshop on Abusive Language Online*, 25-35. Florencia, Italia. doi:10.18653/v1/W19-3504

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (5 de 2017). Automated Hate Speech Detection and the Problem of Offensive Language. *11*, 512-515. Obtenido de https://ojs.aaai.org/index.php/ICWSM/article/view/14955

Fersini, E., Nozza, D., & Rosso, P. (2020). AMI @ EVALITA2020: Automatic Misogyny Identification. doi:10.4000/books.aaccademia.6764

Fersini, E., Rosso, P., & Anzovino, M. (2018). Overview of the Task on Automatic Misogyny Identification at IberEval 2018.

Fortuna, P., Silva, J. R., Soler-Company, J., Wanner, L., & Nunes, S. (2019). A Hierarchically-Labeled Portuguese Hate Speech Dataset. *Proceedings of the 3rd Workshop on Abusive Language Online (ALW3)*.

Founta, A.-M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., . . . Kourtellis, N. (2018). Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. Obtenido de http://arxiv.org/abs/1802.00393

Frenda, S., Ghanem, B., Montes, M., & Rosso, P. (2019). Online Hate Speech against Women: Automatic Identification of Misogyny and Sexism on Twitter. *36*, 4743-4752. doi:10.3233/JIFS-179023

Fulper, R., Ciampaglia, G. L., Ferrara, E., Menczer, F., Ahn, Y., Flammini, A., . . . Rowe, K. (6 de 2015). Misogynistic Language on Twitter and Sexual Violence. En *Proc. ACM Web Science Workshop on Computational Approaches to Social Modeling (ChASM), 2014*. doi:10.6084/m9.figshare.1291081

García-Díaz, J. A., Cánovas-García, M., Colomo-Palacios, R., & Valencia-García, R. (2021). *Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings* (Vol. 114). doi:10.1016/j.future.2020.08.032.

García-Díaz, J. A., Colomo-Palacios, R., & Valencia-García, R. (2021). UMUTeam at EXIST 2021, Sexist Language Identification based on Linguistic Features and Transformers in Spanish and English. Obtenido de https://hdl.handle.net/11250/2831044

Gibert, O. d., Perez, N., García-Pablos, A., & Cuadros, M. (10 de 2018). Hate Speech Dataset from a White Supremacy Forum. *Proceedings of the 2nd Workshop on Abusive Language Online ({ALW}2)*, 11-20. Bruselas, Bélgica. doi:10.18653/v1/W18-5102

Goenaga, I., Atutxa, A., Gojenola, K., Casillas, A., Ilarraza, A. D., Ezeiza, N., . . . Viñaspre, O. P. (2018). Automatic Misogyny Identification Using Neural Networks.

Goled, S. (17 de 3 de 2021). Why Transformers are increasingly becoming as important As RNN and CNN? *Analyticsindiamag*.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., . . . Pang, R. (2020). Conformer: Convolution-augmented Transformer for Speech Recognition. *Proc. Interspeech 2020*, 5036-5040. doi:10.21437/Interspeech.2020-3015

Hajarian, M., & Khanbabaloo, Z. (2021). Toward Stopping Incel Rebellion: Detecting Incels in Social Media Using Sentiment Analysis. *2021 7th International Conference on Web Research (ICWR)*, 169-174. doi:10.1109/ICWR51868.2021.9443027

Kennedy, B., Atari, M., Davani, A. M., Yeh, L., Omrani, A., Kim, Y., . . . al., e. (4 de 2022). The Gab Hate Corpus. doi:10.17605/OSF.IO/EDUA3

Kumar, R., Pal, S., & Pamula, R. (2021). Sexism Detection in English and Spanish Tweets.

Leite, J. A., Silva, D. F., Bontcheva, K., & Scarton, C. (2020). Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis. Obtenido de https://arxiv.org/abs/2010.04543

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2021). A Survey of Transformers. *abs/2106.04554*. CoRR. Obtenido de https://arxiv.org/abs/2106.04554

Metz, C. (24 de 11 de 2020). Meet GPT-3. It Has Learned to Code (and Blog and Argue). *The New York Times*.

Park, J. H., & Fung, P. (2017). One-step and Two-step Classification for Abusive Language Detection on Twitter. Obtenido de http://arxiv.org/abs/1706.01206

Pascoe, C. J., & Diefendorf, S. (2019). No Homo: Gendered Dimensions of Homophobic Epithets Online. doi:10.1007/s11199-018-0926-4

Paula, A. F., Silva, R. F., & Schlicht, I. B. (2021). Sexism Prediction in Spanish and English Tweets Using Monolingual and Multilingual BERT and Ensemble Models.

Pelle, R. d., & Moreira, V. (2017). Offensive Comments in the Brazilian Web: a dataset and baseline results. *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. São Paulo. doi:10.5753/brasnam.2017.3260

Plaza-Del-Arco, F.-M., Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2020). *Detecting Misogyny and Xenophobia in Spanish Tweets Using Language Technologies* (Vol. 20). New York, USA: ACM Trans. Internet Technol. doi:10.1145/3369869

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. doi:10.1007/s11431-020-1647-3

Rodríguez-Sánchez, F., Carrillo-de-Albornoz, J., Plaza Morales, L., Gonzalo Arroyo, J., Rosso, P., Comet, M., & Donoso, T. (2021). Overview of EXIST 2021: sEXism Identification in Social neTworks. doi:10.26342/2021-67-17

Sanguinetti, M., Comandini, G., Nuovo, E., Frenda, S., Stranisci, M., Bosco, C., . . . Russo, I. (12 de 2020). HaSpeeDe 2 @ EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task.

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., & Stranisci, M. (2018). An Italian Twitter Corpus of Hate Speech against Immigrants. *Proceedings of the 11th Conference on Language Resources and Evaluation*, 2798-2895. Miyazaki, Japón.

Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., & Lee, A. A. (2019). Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *5*, 1572-1583. ACS Central Science. doi:10.1021/acscentsci.9b00576

Shushkevich, E., & Cardiff, J. (2019). *Automatic Misogyny Detection in Social Media: A Survey* (Vol. 23). Ciudad de México: scielomx. doi:10.13053/cys-23-4-3299

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention Is All You Need. CoRR. Obtenido de http://arxiv.org/abs/1706.03762

Vera Lagos, V. (10 de 2021). Detección de misoginia en textos cortos mediante clasificadores supervisados. Puebla, México. Obtenido de https://hdl.handle.net/20.500.12371/15437

Waseem, Z. (2016). Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. *Proceedings of the First Workshop on NLP and Computational Social Science*, 138-142. Austin, Texas: Association for Computational Linguistics. Obtenido de http://aclweb.org/anthology/W16-5618

Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. *Proceedings of the NAACL Student Research Workshop*, 88-93. San Diego, California: Association for Computational Linguistics. Obtenido de http://www.aclweb.org/anthology/N16-2013

Zhang, Z., & Luo, L. (2018). Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter. Obtenido de http://arxiv.org/abs/1803.03662

APPENDIXES

Appendix A. Glossary

This appendix consists of short descriptions of some terms and expressions used during this document.

- **Incel:** Short for "involuntary celibate", incel refers to, usually, a young man that sees himself as unable to be intimate with women because of his physical appearance. Incels are usually hostile towards themselves and women, blaming them for not being attracted to them. This community developed online on forums such as Reddit and 4chan.

- **Men's Rights Activists:** Members of the Men's Rights Movement, an anti-feminist group that discuss topics in defence of men and mainly supported by the "alt-right", a far-right white nationalist movement.

- **Retweet**:

Appendix B. Hate speech

In this appendix I will briefly explain what I determined to be hate speech in this project.

Finding a robust definition of hate speech is a complicated task; there is no formal definition for it in International Human Rights Law, and every country has different laws when it comes to hate speech. However, the definition we find when searching for hate speech is "abusive or threatening speech or writing that expresses prejudice against a particular group, especially on the basis of race, religion, or sexual orientation". Knowing this, I wish to note that often hate speech is intertwined with social prejudice; and therefore, a comment making fun of a white person's race isn't the same as one vilifying dark skin, since white people haven't been consistently oppressed. The case is the same with any other social division (religion, gender identity, etc.)

Taking this into consideration, I will disregard hateful comments targeted towards privileged groups classified as hate speech. These comments are definitely offensive, hurtful and unnecessary, but are not supported by the oppressive backbone of today's society.

Appendix C. Bias in NLP

In this appendix I'll be explain what I believe is a very important point on data annotation. In the figure below we can see the sex, sexual orientation and ethnicity of the annotators of the ToLD-Br dataset seen in section 3.12. As we can see, the main ethnicity is white and the main sexual orientation is heterosexual. This means that there is a bias in this dataset. This doesn't mean that the data is not trustworthy, it simply means that it's necessary to take this into consideration when working with manually labelled data.

	Categories	# annotators
Sex	Male	18
	Female	24
Sexual orientation	Heterosexual	22
	Bisexual	12
	Homosexual	5
	Pansexual	3
Ethnicity	White	25
	Brown	9
	Black	5
	Asian	2
	Non-Declared	1

Figure 3. ToLD-Br annotator demographic (Leite, Silva, Bontcheva, & Scarton, 2020)

Even though not all of the used datasets have specific information about the annotators, It's safe to say the data would still have biases.

LIST OF BADY BIASED PAPERS MENTIONED IN (Blodgett, Barocas, III, & Wallach, 2020): (check out golbeck et al?) mentioned in large data English founta

- (Davidson, Bhattacharya, & Weber, Racial Bias in Hate Speech and Abusive Language Detection Datasets, 2019) (check out aave thing)