Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2362–2370
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Hate Speech Dynamics Against African descent, Roma and LGBTQ+ Communities in Portugal

Paula Carvalho[1], **Bernardo Matos**[1,2], **Raquel Santos**[1,2], **Fernando Batista**[1,3], **Ricardo Ribeiro**[1,3]
[1] INESC-ID Lisboa, Rua Alves Redol, 9, 1000-029 Lisboa, Portugal
[2] Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal
[3] Instituto Universitário de Lisboa (ISCTE-IUL), Lisboa, Portugal
pcc@inesc-id.pt, {bernando.cunha.matos, raquel.bento.santos}@tecnico.ulisboa.pt
{fernando.batista, ricardo.ribeiro}@inesc-id.pt

Abstract

This paper introduces FIGHT, a dataset containing 63,450 tweets, posted before and after the official declaration of Covid-19 as a pandemic by online users in Portugal. This resource aims at contributing to the analysis of online hate speech targeting the most representative minorities in Portugal, namely the African descent and the Roma communities, and the LGBTQ+ community, the most commonly reported target of hate speech in social media at the European context. We present the methods for collecting the data, and provide insightful statistics on the distribution of tweets included in FIGHT, considering both the temporal and spatial dimensions. We also analyze the availability over time of tweets targeting the aforementioned communities, distinguishing public, private, and deleted tweets. We believe this study will contribute to better understand the dynamics of online hate speech in Portugal, particularly in adverse contexts, such as a pandemic outbreak, allowing the development of more informed and accurate hate speech resources for Portuguese.

Keywords: hate speech, offensive speech, Twitter

1. Introduction

Most research in hate speech detection focuses on the creation of language resources, and on the development of methods and tools for automatically detecting offensive and abusive language (Fortuna and Nunes, 2018). However, the lack of consensus on the definition and characterization of hate speech has led to the creation of heterogeneous resources, particularly annotated corpora, making it difficult to compare them (Poletto et al., 2021). In addition, hate speech is intrinsically dependent on the sociocultural context, which means that existing resources cannot be directly transferable or easily adaptable to other linguistic and pragmatic contexts (Nozza, 2021). Although there are few corpora specifically designed for detecting hate speech in Portuguese (Fortuna et al., 2019), their usefulness is quite limited to study spatiotemporally delimited phenomena, such as the dynamics of online hate speech in Portugal, before and during the Covid-19 pandemic. Moreover, the data included in the existing corpora is often selected based on generic lexical-based approaches, using closed lists of keywords with negative polarity, typically involving epithets and slurs that may be used to incite hatred or violence against an individual or a group. Used in isolation, this selection method leaves out an immense set of potentially relevant hatred content, including indirect or covert hate speech (Baider and Constantinou, 2020; Kumar et al., 2018), often resorting to rhetorical figures, such as irony, sarcasm, humor, analogy, metaphor, and rhetorical questions, and then preventing an in-depth understanding of the nature and extent of this phenomenon.

To address the research gaps mentioned above, we created FIGHT (**FI**ndin**G** **H**ate Speech in **T**witter), a dataset containing 63,450 Portuguese tweets, posted by 6,728 different users located in Portugal, as defined in their profile account. The selected tweets cover about 1,5 years before and 1,5 years after the official declaration of Covid-19 as pandemic by the World Health Organization (WHO). In particular, this corpus aims at contributing to the analysis of hate speech by the Portuguese online community targeting the most representative minorities in Portugal, namely the African descent and the Roma communities (Maeso, 2021). Moreover, since the LGBTQ+ community is still the most commonly reported target of hate speech in social media at the European context (Wigand et al., 2021), we have also decided to include it in our study.

FIGHT is composed of (i) tweets mentioning the above-mentioned target groups, and (ii) tweets potentially conveying offensive or hate speech against those groups. While the former will be an important source to further investigate either indirect hate speech or counterspeech (Benesch et al., 2016; Chung et al., 2019), especially by exploring the conversations associated with the collected tweets, the latter will allow investigating potential instances of direct hate speech, and contrasting them with offensive speech and impoliteness (Culpeper, 2021).

This paper presents the methods underlying the data collection, and provides some statistics on the distribution of tweets included in FIGHT, taking into account both the temporal and spatial dimensions. Furthermore, we present some statistics on the availability over time

of the tweets for each class, distinguishing between public, private and deleted tweets. The information on deleted tweets could be particularly relevant to further investigate the relationship between this action and the potential hatred conveyed in messages (Bhattacharya and Ganguly, 2016).

The results achieved in an initial manual annotation trial, based on a small sample from FIGHT, suggest that 40% of the tweets pre-classified as conveying potentially offensive or hate speech are effectively offensive or hateful. We intend to fully annotate this corpus in near future, based on solid guidelines being created by the project's team, to consolidate the results achieved.

We believe this study will contribute to better understand the dynamics of online hate speech in Portuguese, particularly in adverse contexts, such as a pandemic outbreak.

The remainder of the paper is structured as follows: Section 2 presents the related work, with a special focus on the hate speech resources available for Portuguese; Section 3 describes the methods we used on creation of FIGHT dataset; the characterization of this language resource is discussed in Section 4; Section 5 presents the results of the annotation trial conducted in this study; and finally Section 6 provides the main conclusions, future directions, and the ethical concerns that must be taken into account.

2. Related Work

Hate speech has been flagged as a serious concern across social media platforms worldwide, which has contributed to the growing interest in developing resources and methods for its automatic detection (Fortuna and Nunes, 2018; Schmidt and Wiegand, 2017; Poletto et al., 2021). In the context of the Covid-19 pandemic, emerging reports show that the corona virus outbreak is related to the increase of discrimination and racist attacks, especially against Chinese people and people with Asian identity features (Ziems et al., 2020). An infodemiological analysis involving the Twitter communities from the United States and Philippines demonstrates that the spread of hate speech around Covid-19 has similar reproduction rates as other Covid-19 information on Twitter (Uyheng and Carley, 2021).

Broadly understood as any kind of communication "that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factor" (United Nations, 2019, p. 2), hate speech is often confused with other instances of offensive language (Davidson et al., 2017) or language aggression (Basile et al., 2019), making its automatic detection harder. In fact, there is not a unique and consensual definition of hate speech, leading to a heterogeneity of language resources specifically tailored to detect this phenomenon. This is often due to the multiplicity of interpretations of the term "hateful", which is often mixed up with other semantically related concepts, such as abusive, toxic, offensive or aggressive language (Poletto et al., 2021).

In this paper, hate speech is operationalized through the following coexisting conditions: (i) hate speech has a specific target that can be mentioned explicitly or implicitly in text; (ii) hate speech targets correspond to vulnerable or historically marginalized groups (usually minority groups) or individuals targeted for belonging to those groups; (iii) hate speech typically spreads or supports hatred, or incites violence against the aforementioned targets, by disparaging, humiliating, discriminating, or even threatening them on the basis of specific identity factors; and (iv) hate speech can be expressed both explicitly (or overly) and implicitly (or covertly).

The heterogeneity among the existing hate speech corpora is also explained by the diversity of target categories and attributes being considered. While some works are mainly concerned with distinguishing generic categories, such as *racism* or *sexism* (Waseem, 2016), others have adopted complex hierarchical labeling schema, including dozens of hate speech categories and subcategories (Fortuna et al., 2019).

Despite there are several resources and benchmark corpora for many different languages, in particular for English (Poletto et al., 2021), we have found only four hate speech datasets for Portuguese.[1] Pelle and Moreira (2017) developed a corpus with 1,250 comments randomly extracted from the most popular Brazilian news website. Those comments focus on political and sports news, whose topics could generate more controversy, and consequently more hate speech. Each comment was classified as being offensive or not offensive, and the former were categorized into one of the following hate speech classes: *xenophobia*, *homophobia*, *sexism*, *racism*, *cursing* and *religious intolerance*. Around 20% of the annotated comments were classified as offensive. Fortuna et al. (2019) have compiled a corpus of 5,668 Portuguese tweets, posted by 115 different users, which were manually classified as conveying hate speech or not; hatred messages were then classified according to its target, following a fine-grained hierarchical multiple label scheme, including 81 hate speech categories. Tweets were retrieved by applying a list of offensive keywords and by selecting the users who usually post hateful comments. Around 22% of the tweets were identified as conveying hate speech. Leite et al. (2020) created a dataset composed by 21,000 Portuguese Brazilian tweets. These posts were retrieved by applying a filtering list of offensive keywords and also by considering keywords related to influential Brazilian users that may be victims of hate speech or abuse. The tweets were assigned with one of the following categories: *LGBTQphobia*, *obscene*, *insult*, *racism*, *misogyny*, and *xenophobia*. According

[1] https://hatespeechdata.com/

to the inter-annotator agreement (IAA) results reported by the authors, *LGBTQphobia* was the most consensual class among the annotators. On the contrary, *obscene* and *racism* classes have achieved the lowest agreement. Lastly, Vargas et al. (2021) present a corpus of 7,000 comments extracted from Instagram posts of six Brazilian political personalities. The messages in the corpus were classified as being offensive or non-offensive; offensive messages were then classified according to the offense's intensity. The targets considered in this corpus were *xenophobia*, *racism*, *homophobia*, *sexism*, *religious intolerance*, *partyism*, *apology to the dictatorship*, *antisemitism* and *fatphobia*. Half of the comments were labeled offensive, and from those 11% were classified as highly offensive, 15% as moderately offensive, and finally 24% as slightly offensive.

Contrarily to the previously described approaches, we did not perform our selection by searching specific topics or actors. Instead, our selection was based on potential mentions to hate speech targets. However, contrarily to the approaches focusing on the targets, we did not restrict our selection to specific individuals or personalities, but to well-founded protected communities. In addition, in spite of using a lexicon of potential offensive terms to search potentially relevant tweets, our approach differ from the others by combining this lexicon with a lexicon describing words and expressions often used to mention the concerned targets, leading to a more refined search. The *target lexicon* was the unique lexicon applied individually, allowing to select a more comprehensive dataset, potentially including indirect hate speech. Finally, our data selection followed specific spatiotemporal criteria, often neglected in data selection, allowing to study the hate speech phenomena within a specific geographic space and time context.

3. Collecting the Data

Since we are interested in analyzing the dynamics of hate speech and related phenomena within the Portuguese context, we decided to first explore an existing database composed of tweets that have been collected daily since 2015. Next, we used the Twitter API to fill potential gaps in data collection. By combining both sources of information, we obtained an updated and robust dataset, which includes the information currently available on Twitter, and information on tweets previously collected that are no longer available, because they were deleted or the Twitter account where they were posted was removed.

For conducting this specific research, we have applied the following selection criteria:

Time span We restricted the data selection to a time span of about 3 years, from August 1, 2018 to October 31, 2021. This time frame allows us to study the potential relationship between the Covid-19 pandemic and the evolution of hate speech in Portuguese social media, particularly in Twitter. The reference date for the beginning of Covid-19 is March 11, 2020, when the Covid-19 outbreak was declared as a pandemic by the World Health Organization (WHO).

Geography To retrieve only the tweets posted by the Portuguese community, we have selected the tweets published by users located in Portugal, according to their Twitter profile information. Those tweets were then assigned with the information on the Portuguese region, based on the Nomenclature of Territorial Units for Statistics (NUTS) II.[2]

Lexicon We have created a lexicon composed of 259 words and expressions often used to mention the targets we are interested in monitoring, particularly the African descent, Roma, and LBGTQI communities. To select the potential targets, we considered only the unambiguous forms associated with each semantic category, corresponding to a total of 174 entries (e.g. *Africans*). In this case, the ambiguous forms, such as *preto* ('black'), were not considered, since they can be used in a variety of contexts with a different meaning (e.g. *Eu adoro esse casaco preto*, 'I love that black coat'). In addition, we have created a lexicon including approx. 800 inflected forms that are often used to insult or offend the previously mentioned targets. This lexicon was combined with the previous one (this time including both ambiguous and unambiguous forms) to retrieve potentially offensive and hateful messages targeting each protected community (e.g. *É o* ***preto*** *mais* ***burro*** *que já vi mano*, 'It's the dumbest nigga I've ever seen bro').

Figure 1 represents the pipeline of data collection approach. Its main components are described as follows:

- Both the Twitter API and the existing database are explored to retrieve: (i) the tweets containing unambiguous terms described in the target lexicon; and (ii) the tweets containing terms from both the target and the offensive lexicons, considering the temporal and geographic constraints previously defined.
- The information on each tweet collected from the database was compared with the information currently available on Twitter. The metrics on publicly available tweets were then updated. The conversations associated with potentially offensive tweets – resulting from combining the above mentioned lexicons – were also extracted, since their content might be extremely relevant to study the expression of hate speech and counterspeech in future research.
- The Twitter API allowed us to complement the existing collection of tweets by adding existing

[2] https://www.pordata.pt/

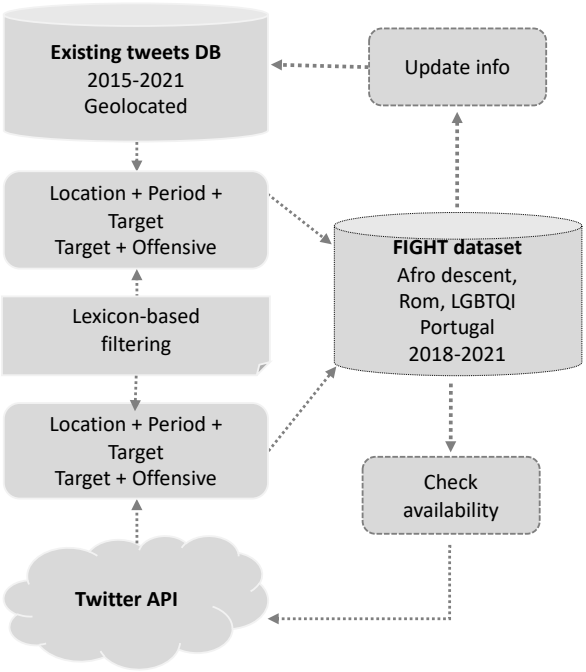

Figure 1: Pipeline for data collection.

Data Source	Target	Off/HS	Total
DB (existing)	35,832	5,576	41,408
Twitter API (new)	17,947	4,095	22,042
Total	**53,779**	**9,671**	**63,450**

Table 1: Distribution of tweets in FIGHT, according to the data source.

tweets there were not already available in the existing DB, due to eventual problems that may have occurred during the data collection process performed on a daily basis over years.

4. FIGHT Data Collection

The FIGHT (**FI**ndin**G** **H**ate in **T**witter) data collection includes the tweets extracted from both the database and Twitter API according to the criteria previously described. The final data collection is composed of 63,450 tweets, posted by 6,728 different users; from those tweets, 41,408 were extracted from the existing database (cf. Table 1). When comparing the data collected in the existing database with the one retrieved by the Twitter API, we found that there were (i) tweets in the database that are no longer available on Twitter, and (ii) tweets available in Twitter that, for some reason, were not present in the database. After merging both data, we obtained 53,779 tweets, from which 22,042 where retrieved from the Twitter API. As described in Table 2, those tweets include the mention to the potential targets covered in this study. The most representative class in our collection is LGBTQ+, followed by the African descent, and finally, with a much fewer representation, the Roma community.

When restricting the search to the combination of a potential mention to the target with a potentially offensive term, we got 9,671 tweets. Interestingly, the African descent is the most representative class in this case (6,678 tweets), followed by the LGBTQ+ (2,647 tweets), and finally the Roma community (346 tweets).

	Before Covid		During Covid	
	Target	**Off/HS**	**Target**	**Off/HS**
Afro	10,886	3,472	12,010	3,206
Roma	1,476	146	1,560	200
LGBTQ+	15,622	1,415	12,245	1,232
Total	**27,984**	**5,033**	**25,815**	**4,638**

Table 2: Distribution of tweets in FIGHT, posted before and after the official declaration of Covid-19 as a pandemic, for each class.

4.1. Tweets Distribution over Time

Figure 2 presents the number of tweets covered in FIGHT, on a monthly basis, considering both the tweets containing a potential mention to each covered target, and those containing potential hate speech (dotted lines). Taking as reference the official declaration of Covid-19 as a pandemic, one can observe that, especially in the last months, the number of tweets targeting the protected groups has been decreased. Similarly, the tweets containing potentially hate speech have also decreased in number compared to the time before the pandemic.

As illustrated in Figure 2, there is a clear relationship between the tweets potentially mentioning the Roma, LGBTQ+, and the African descent communities, and the tweets potentially containing offensive or hate speech against each aforementioned target. In fact, the general trends and major peaks seem to be related in both collections. Regarding the classes' distribution, tweets mentioning the LGBTQ+ community – one of the most commonly reported target of hate speech in social media at the European context – have generally been the most prevalent over time. However, at least in FIGHT, this trend seems to be attenuating, especially in recent months.

Particularly concerning the potential offensive and hateful tweets, no abrupt peaks are observed for the LGBTQ+ community. In this case, the most prevalent class over time is the African descent community, one of the most representative minorities in Portugal.

The highest peaks occur in the months next to the declaration of the Covid-19 pandemic, with particular emphasis for the African descent community. The interpretation of these values requires the inspection of potential events that may influence the users' activity in social networks.

With respect to African descent community, the first pick of tweets was in February 2020. This may be related with an event involving an African descent foot-

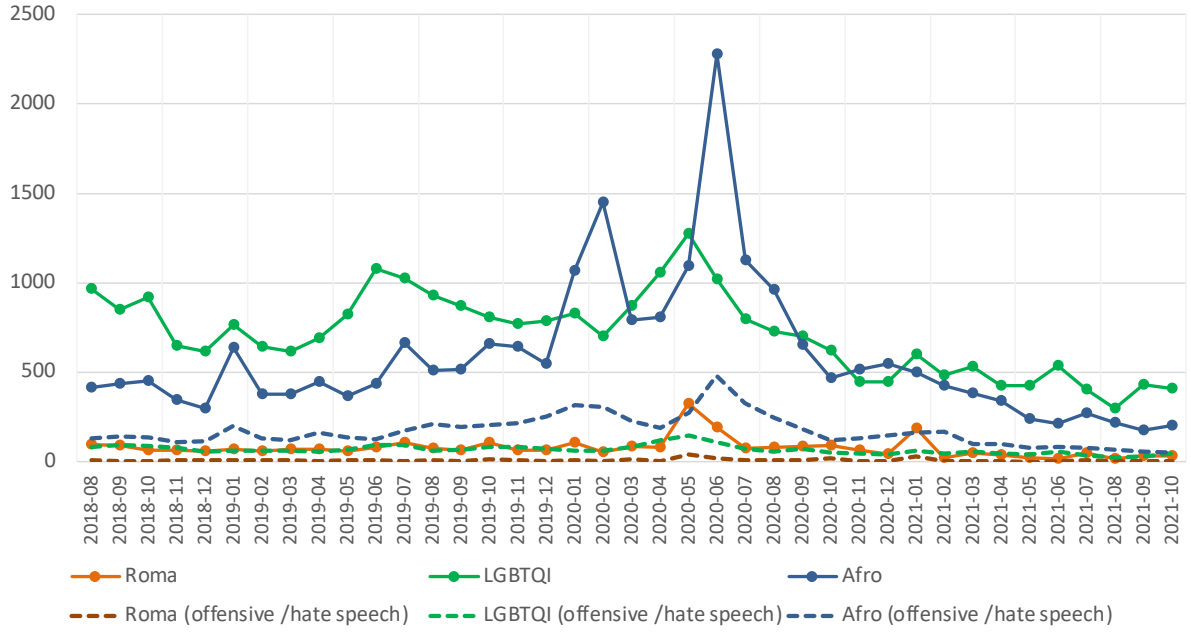

Figure 2: Number of monthly tweets mentioning each potential target, and the corresponding portion of tweets potentially conveying offensive or hate speech.

ball player, who abandoned a match due to the racist insults and slurs he experienced from a group of football fans. This incident stimulated the discussion about racism in either conventional or social media. The second and the highest peak occurred in June 2020 seems to be directly related to the murder of George Floyd in May 2020 and the consequent anti-racism protests, particularly in Portugal.

Regarding the expression of potential hate speech targeting the Roma community, we can observe a peak in May 2020, which might be related to the proposal of a special confinement plan for the Roma community in Portugal, made by a Portuguese deputy who is also the president of a national conservative, right-wing political party in Portugal. This lead to a public debate involving multiple figures, repudiating or supporting the discrimination against this community. Another peak occurs in January 2021, during the period of the Portuguese presidential elections, in which the same political actor ran.

4.2. Tweets Distribution per Region

Figure 3 presents the proportion of potentially offensive or hatred tweets by target and by region, before and during the Covid-19 pandemic, taking into consideration the total number of collected tweets (about 15 million) by the Portuguese community in Twitter during the period under analysis. The most representative target group in both periods of time is the African descent community, who stands out from the remaining classes in all regions, with the exception of Madeira, whose representativeness is close to the LGBTQ+ community. In fact, before the Covid-19 pandemic the most envisaged group in this region was the LGBTQ+ community, but this has changed in the period during the Covid-19, following the national trend. Proportionally, the regions of Alentejo and Azores have the highest number of hateful messages against the African descent community, especially before the pandemic. The most regular behavior along the period considered for all hatred groups is observed in the regions of Lisbon and the North of Portugal. In terms of evolution, it is important to highlight the increase of potentially offensive or hateful messages targeting the African descent community in Algarve, during the Covid-19, contradicting the downward trend at national level. Particularly regarding the messages targeting the Roma community, it is recorded a slight increase in Madeira and in Alentejo, where several conflicts with this community have been reported in the recent times.

Table 3 shows the 20 content words (i.e. nouns, adjectives, verbs, and adverbs) most frequent in the retrieved tweets for the most representative regions (considering the number of retrieved tweets) in Portugal. Interestingly, the most frequent terms in tweets from all regions include words related to the LGBTQ+ community (e.g., *gay* and *paneleiro*; 'gay' offensive), reinforcing the trend previously reported. The mentions to African descent community (e.g., *preto* and *angolano*; 'Black' and 'Angolan') are also highly frequent in all regions. Explicit references to the Roma community are mostly frequent in Central Portugal, and in the regions of Lisbon, and Alentejo (e.g., *cigano*; 'Roma' or 'gypsy'). Moreover, it is important to stress the huge

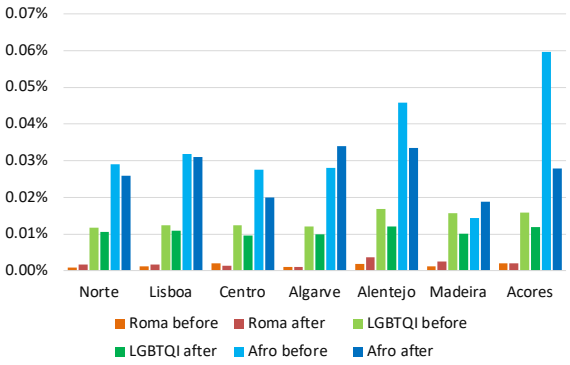

Figure 3: Proportion of tweets containing potentially offensive or hate speech, by online users from each Portuguese region, before and during Covid-19.

prevalence of terms like *racista*, and *racismo* ('racist' and 'racism'), in all the collections, clearly related to the topic approached in the paper.

The frequent use of offensive, pejorative, and insulting words (e.g., *merda* and *caralho*; 'shit' and 'fuck') supports the idea that the tweets included in FIGHT could effectively be an important source of offensive and hate speech.

Particularly regarding adverbs, it must be stressed the use of the negative adverbs *não* ('no') and *nem* ('neither') in tweets from all regions; intensifiers are also highly frequent (e.g., *muito* and *mesmo*; 'very' and 'really'), which might suggest that the tweets in FIGHT convey strongly marked sentiment and opinions.

4.3. Public, Private, and Deleted Tweets

Since the tweets retrieved from the existing database were collected on the date they were published, it is possible to identify the tweets that were deleted later on either by their owner or suspended by Twitter, in case of violating Twitter's hateful conduct policy.[3] Figure 4 presents, for each target, the monthly number of tweets potentially containing hate speech that are not currently available on Twitter, either because they were deleted, or marked as private. The blue bars correspond to tweets currently available using the Twitter API. The yellow bars correspond to tweets that still exist, but the API could not retrieve them, since their authors have a private account. The orange bars correspond to deleted tweets. For those, we distinguish between the deleted tweets whose the author's account still exists, and the ones whose the Twitter's account was removed.

The number of restricted or deleted tweets increases together with the time span. Looking back at these data in near future is crucial to understand whether the tweets' deletion is more or less immediate (Almuhimedi et al., 2013) or, on the contrary, it is primarily reflected later.

[3] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

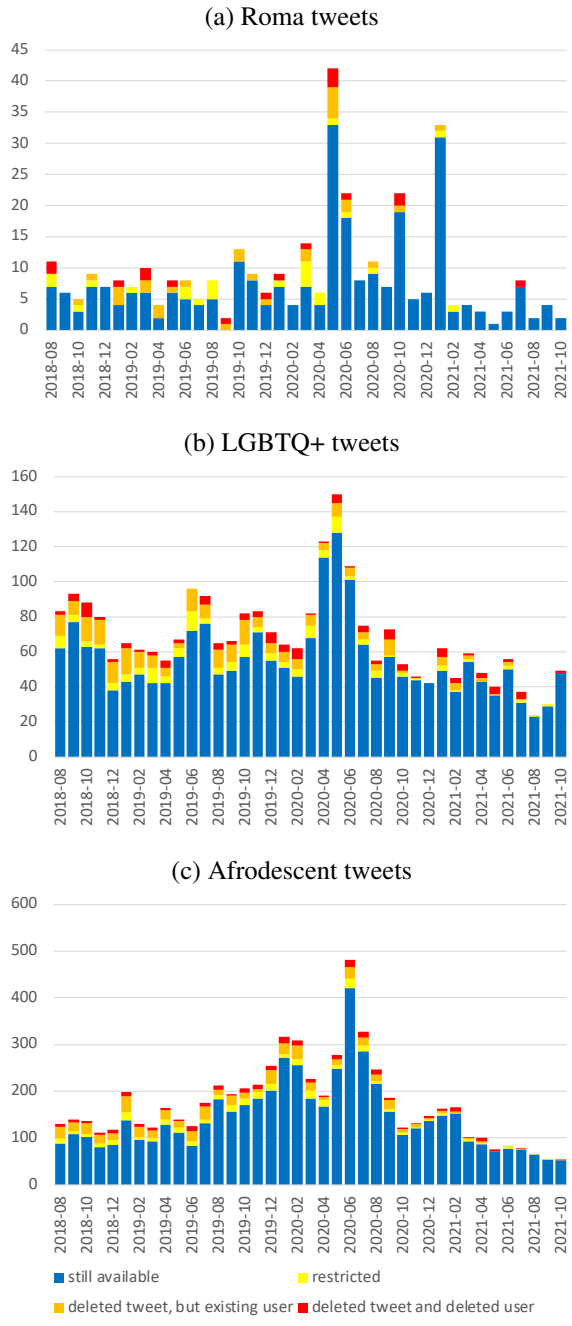

Figure 4: Proportion of public, private, and deleted tweets potentially conveying hate speech against the (a) Roma, (b) LGBTQ+, and (c) African descent communities.

Table 4 presents the overall distribution of tweets for each FIGHT category that are no longer available, distinguishing those that were kept private by their owners and the ones that were permanently deleted. As one can observe, in our collection, deleted tweets correspond to more than a double of private tweets. Furthermore, the highest percentage of deleted tweets seem to be related with the Roma community, followed by the LGBTI community, either considering the messages that in-

Norte		Centro		Lisboa		Alentejo		Algarve	
2974	gay	861	gay	3587	gay	968	não	861	gay
2211	não	792	não	2568	não	763	gay	792	não
1427	racista	475	racista	2182	racista	479	racista	475	racista
1335	racismo	302	racismo	1821	racismo	441	racismo	302	racismo
961	paneleiro	255	mais	1101	pessoa	279	merda	255	mais
720	pessoa	212	são	1081	quando	266	cigano	212	são
668	quando	205	cigano	995	mesmo	264	ser	205	cigano
609	muito	201	paneleiro	988	cigano	257	são	201	paneleiro
638	mesmo	190	pessoa	948	muito	251	pessoa	190	pessoa
609	muito	183	ser	936	angolano	242	paneleiro	183	ser
606	agora	179	preto	930	preto	219	preto	179	preto
588	merda	166	merda	888	foi	207	muito	166	merda
570	são	162	muito	887	são	202	foi	162	muito
564	ser	156	portugal	861	merda	200	está	156	portugal
545	caralho	148	mesmo	845	branco	191	nem	148	mesmo
543	preto	144	foi	844	agora	184	vai	144	foi
499	vai	142	ainda	824	paneleiro	179	tem	142	ainda
475	sempre	136	quando	810	nem	177	mesmo	136	quando
468	branco	135	agora	763	bem	169	agora	135	agora
449	sim	133	nem	725	tem	162	quando	133	nem

Table 3: List of the 20 most frequent content words, considering the most representative Portuguese regions (in terms of retrieved tweets) in FIGHT. Each word is preceded by its frequency.

Class	Target		Off/HS	
	Priv	Del	Priv	Del
Afrodescent	5.43	16.62	5.77	13.85
Roma	7.71	19.43	8.27	16.19
LGBTQ+	6.40	17.08	6.19	16.07

Table 4: Percentage tweets in FIGHT that are currently unavailable, because they are kept private (Priv) or deleted (Del), considering both lexical criteria used.

clude at least an unambiguous mention to the potential target, or the messages potentially conveying offensive or hate speech (i.e., including a mention to a potential target and an expression of offense or insult). The Pearson correlation coefficient indicates that there is no correlation between the number of published tweets and deleted tweets over time for any of the classes considered ($\rho \leq .01$).

5. Annotation Trial

To assess the data usefulness and reliability, we have randomly selected a data sample of 300 tweets (100 from each target group) classified in FIGHT as potentially containing offensive or hate speech against the protected communities considered in this study. Those tweets were manually annotated by two Master students making part of the project's team, and who are currently developing research on automatic hate speech detection. The annotators were asked to identify whether the tweet message (i) conveys hate speech; (ii) is offensive; (iii) is ambiguous, vague or unclear; or finally (iv) is non-relevant, because it contains neither offensive nor hate speech. The annotators were provided with detailed guidelines, developed in the scope of this project, allowing to clearly distinguish offensive from hate speech, which are often mixed in literature. According to the aforementioned guidelines, Examples 1 and 2 should be analyzed as conveying hate speech. On the contrary, despite being offensive or insulting, Examples 3 and 4 should not be considered hateful, since they do not attack an individual or a group on the basis of their identity characteristics.

1. *Racismo o c@ralho! se não fossem esses parasitas da sociedade que não querem fazer nada, Portugal era um paraíso.* 'Fuck the racism! If it were not those social parasites that don't want to do anything, Portugal was a paradise.'

2. *Coitadinhos dos "feirantes", vão ficar sem os benefícios.* 'Poor "marketeers" [reference to Roma], they will lose their benefits.'

3. *Deve ter nascido num pardieiro.*
'You were certainly born in a dump.'

4. *É tudo a mesma bosta, todos esses vermes são racistas e xenofóbicos.* 'It's all the same crap, all these worms are racist and xenophobic.'

In average, 40% of the tweets composing the data sample actually contain offensive or hate speech. This supports our data collection strategy, given the difficulty in identifying data representing such diffused phenomena in social platforms like Twitter. To evaluate the annotations reliability, we then conducted an inter-agreement (IAA) study, using Krippendorff's alpha coefficient (Krippendorff, 2004). Table 5 presents

Variables	IAA	Agree	Disagree
Hate Speech	**0.752**	283	17
Offensive Speech	0.646	259	41
Unclear	0.237	260	40
Non-Relevant	0.726	259	41

Table 5: IAA results for a sample of 300 tweets, measured by Krippendorff's alpha coefficient. The table also presents the total number of agreements (Agree) and disagreements (Disagree) between the annotators.

the agreement obtained for each variable considered in the annotation study. With the exception of the variable *unclear*, which had poor to fair agreement, all the remaining variables achieved a substantial agreement. Interestingly, the highest agreement achieved concerns hate speech detection; in comparison, offensive speech seems more difficult to recognize. In spite of being promising, any type of generalization based on such a limited set of data is incautious. Hence, we intend to further substantially enlarge the annotated dataset, and potentially involving more annotators, to validate the results achieved.

6. Main Conclusions and Future Directions

We have presented FIGHT, a dataset containing 63,450 tweets, covering about a year and a half before and after the official declaration of Covid-19, posted by online users located in Portugal, according to their profile account. This language resource is focused on the most representative minorities in Portugal, namely the African descent and the Roma communities, and the LGBTQ+ community, the most commonly reported target of hate speech in social media at the European context. We have presented the methods used for collecting the data, and provided statistics on the distribution of tweets included in FIGHT, taking into account both the temporal and spatial dimensions. We have also presented statistics on the tweets' availability over time, distinguishing public, private and deleted tweets, based on FIGHT.

Overall, the inspection of FIGHT suggests a descending trend in potentially offensive and hate speech on Twitter, particularly in recent months. This may be directly related with the European Union efforts on countering hate speech (European Union, 2016), and the rigorous hateful conduct policies being adopted by social platforms, in particular Twitter.[4] Nevertheless, the general examination of tweets in FIGHT shows that some of them clearly violate those policies.

The data also suggests the highest peaks of tweets are intimately related with controversy events directly or indirectly involving the targets considered in FIGHT. In terms of representativeness, the most prevalent target in our dataset is the African descent group, who also gathered the highest number of potential hatred messages, in all the Portuguese regions over the time period considered. This number is impressive in the regions of Azores and Alentejo before the Covid-19 pandemics; however, it has been decreased significantly during the Covid-19 outbreak. The LGBTQ+ community is the most regularly mentioned target in FIGHT, although the number of potential offensive or hate speech targeting this group is lower than the one targeting the African descent group. Compared to other groups, the Roma community is the least represented in our data collection. Since we used a lexicon-based approach to select both the targets considered in our dataset and the potential hatred messages involving those targets, any type of data generalization is imprudent. Moreover, given we have restricted our selection to tweets geolocated in Portugal, we are aware that FIGHT covers only a very small percentage of data published by the Portuguese online community. Hence, the results presented in the paper should rather be interpreted as important clues to perform a further in-depth investigation, namely by exploring the variables considered in our research.

Concerning future work, the results achieved with our initial annotation experiment, based on a small sample from FIGHT, suggest that 40% of the tweets preclassified in the corpus as conveying potentially offensive or hate speech are effectively offensive or hateful. We intend to fully annotate this corpus in the near future, based on solid guidelines being created by the project's team. In addition, we intend to explore the conversations associated with the collected tweets, to overcome the lexicon-based approach's drawbacks.

We believe this study will contribute to better understand the dynamics of online hate speech in Portuguese, particularly in adverse contexts, such as a pandemic outbreak, and will be an important resource to promote the research of hate speech detection in this language.

The FIGHT data collection follows the Twitter Developer policy,[5] and several procedures must be considered before publishing it. Concerning the publicly available tweets, only the ID will be distributed, together with the respective semi-automatic annotations. Concerning the deleted tweets, their possible inclusion in the FIGHT public version implies anonymization procedures in order to prevent the identification of the user.

7. Acknowledgments

This research was supported by Fundação para a Ciência e a Tecnologia, through the Projects HATE Covid-19 (Proj.759274510), MIMU (FCT PTDC/CCI-CIF/32607/2017), and UIDB/50021/2020.

[4] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

[5] https://developer.twitter.com/en/developer-terms/policy

8. References

Almuhimedi, H., Wilson, S., Liu, B., Sadeh, N., and Acquisti, A. (2013). Tweets are forever: a large-scale quantitative analysis of deleted tweets. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 897–908.

Baider, F. and Constantinou, M. (2020). Covert hate speech: A contrastive study of greek and greek cypriot online discussions with an emphasis on irony. *Journal of Language Aggression and Conflict*, 8(2):262–287.

Basile, V., Bosco, C., Fersini, E., Debora, N., Patti, V., Pardo, F. M. R., Rosso, P., Sanguinetti, M., et al. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics.

Benesch, S., Ruths, D., Dillon, K. P., Saleem, H. M., and Wright, L. (2016). Counterspeech on twitter: A field study. dangerous speech project.

Bhattacharya, P. and Ganguly, N. (2016). Characterizing deleted tweets and their authors. In *Tenth international AAAI conference on web and social media*.

Chung, Y.-L., Kuzmenko, E., Tekiroglu, S. S., and Guerini, M. (2019). Conan–counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. *arXiv preprint arXiv:1910.03270*.

Culpeper, J. (2021). Impoliteness and hate speech: Compare and contrast. *Journal of Pragmatics*.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515, May.

European Union. (2016). The EU Code of Conduct on countering illegal hate speech online. https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en.

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

Fortuna, P., da Silva, J. R., Wanner, L., Nunes, S., et al. (2019). A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104.

Krippendorff, K. (2004). Measuring the reliability of qualitative text analysis data. *Quality and quantity*, 38:787–800.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2018). Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11.

Leite, J. A., Silva, D. F., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. *arXiv preprint arXiv:2010.04543*.

Maeso, S. R. (2021). *O Estado do Racismo em Portugal*. Tinta da China, Lisbon, Portugal, 1st edition.

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

Pelle, R. P. and Moreira, V. P. (2017). Offensive comments in the brazilian web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. SBC.

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

Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

United Nations. (2019). United nations strategy and plan of action on hate speech. Technical report, United Nation.

Uyheng, J. and Carley, K. M. (2021). Characterizing network dynamics of online hate communities around the covid-19 pandemic. *Applied Network Science*, 6(1):1–21.

Vargas, F. A., Carvalho, I., de Góes, F. R., Benevenuto, F., and Pardo, T. A. S. (2021). Building an expert annotated corpus of brazilian instagram comments for hate speech and offensive language detection. *arXiv preprint arXiv:2103.14972*.

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

Wigand, C., Kolanko, K., and Ferroli, J. (2021). 6th evaluation of the EU Code of Conduct. Technical report, European Commission.

Ziems, C., He, B., Soni, S., and Kumar, S. (2020). Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis. *arXiv preprint arXiv:2005.12423*.