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, Rua Alves Redol, 9, 1000-029 Lisboa, Portugal
[2] Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal
[3] Iscte - Instituto Universitário de Lisboa, Lisboa, Portugal
pcc@inesc-id.pt, {bernando.cunha.matos, raquel.bento.santos}@tecnico.ulisboa.pt {fmmb, rdmr}@inesc-id.pt

1 Motivation and goals

Motivation

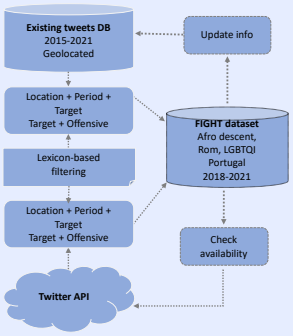
- Heterogenous Hate Speech (HS) resources.
- Few corpora specifically designed for studying HS in Portuguese.
- Spatial and temporal dimensions are not usually considered.

Goals

- Investigate the dynamics of online HS in Portugal, particularly against the most representative marginalized groups.
- Understand the impact of the Covid-19 pandemics on online HS.

2 Collecting the Data

- **Time Span**: August 1, 2018 - October 31, 2021
- **Geography**: Tweets posted by the Portuguese online community
- **Lexical criteria**:
  - Lexicon of 259 words and expressions often used to **mention the targets**
  - Lexicon of 800 words and expressions often used to **insult or offend** the targets



FIGHT - FIndinG Hate in Twitter
63,450 tweets, by 6,728 different users, focused on the African descent, Roma and LGBTQ+ communities

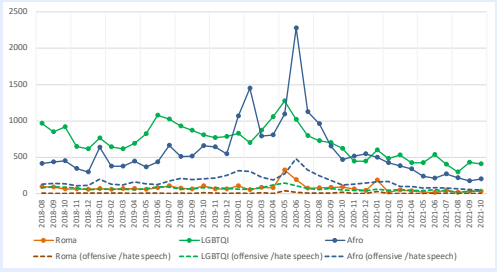
Data Source	Target	Off/HS	Total
DB (existing)	35,832	5,576	41,408
Twitter API (new)	17,947	4,095	22,042
Total	**53,779**	**9,671**	**63,450**

	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**

3 FIGHT Data Collection

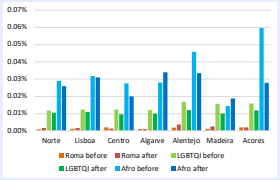
Tweets Distribution over Time

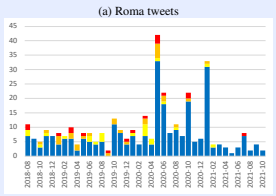
- The highest peaks occur in the months next to the declaration of the Covid-19 pandemic, with particular emphasis for the African descent community.
- The highest peaks of tweets are intimately related with controversy events directly or indirectly involving the target communities considered.

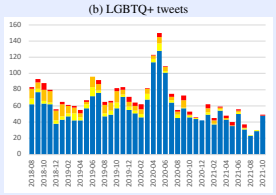


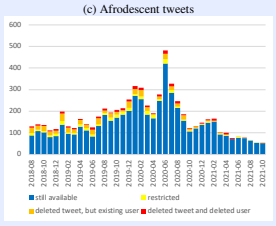
Tweets Distribution per region

- Afro-descendants: the most target group before and after the pandemic
  - Alentejo and Azores: the highest number of hateful messages against this community.
  - Algarve and Madeira: increase of potential HS against this community
- Lisbon and North of Portugal: the most regular behavior.
- Alentejo and Madeira: slight increase of potential HS against Roma.









Public, Private, and Deleted Tweets

- Deleted tweets correspond to more than a double of private tweets.
- There is no correlation between the number of published tweets and deleted tweets over time ($\rho \leq .01$).

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

4 Annotation Trial

- Randomly selected a data sample of 300 tweets (100 from each target group) classified as potentially containing offensive or hate speech.
- Two annotators were asked to identify whether the tweet message (i) conveys hate speech; (ii) is offensive; (iii) is ambiguous, vague or unclear; or (iv) is non-relevant.
- 40% of analysed tweets classified as conveying offensive or hate speech.

Inter-Annotator Agreement (IAA)

- Substantial agreement for hate speech detection (Krippendorff's alpha 0.752)
- Offensive speech seems more difficult to recognize (Krippendorff's alpha 0.646)

5 Conclusions and future work

Conclusions

- Descending trend in potentially offensive and hate speech on Twitter.
- The most prevalent target in FIGHT 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.
- 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.

Future work

- Fully annotate the corpus, based on solid guidelines developed by the project's team (in progress).
- Explore the conversations associated with the collected tweets, to overcome the lexicon-based approach's drawbacks.
- Made the annotated corpus available.

Research supported by FCT, through projects Hate Covid-19 (Proj.759274510), MIMU (FCT PTDC/CCICIF/32607/2017), and UIDB/50021/2020