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.

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 (p ≤ .01).

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

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)

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.