Hate speech detection in Italian tweets

Group 4

November 6, 2024

1 MOTIVATION

Although hate speech is a phenomenon that existed long before the advent of the Web, communication via the Internet has allowed this phenomenon to spread and develop with peculiar characteristics tied to the nature of online media, such as anonymity on one side, and velocity and breadth of proliferation on the other. Given the relevance and dangerous potential of this social phenomenon, there is widespread interest in recognizing and detecting hate speech online, also on an institutional level.

In May 2016 the European Commission, citing the chilling effect of hate speech on the democratic discourse on online platforms, agreed with Facebook, Microsoft, Twitter and YouTube on a "Code of conduct on countering illegal hate speech online" [2]. In this document the major IT Companies made a public commitment to:

1. Have in place clear and effective processes to review notifications regarding hate speech on their platforms;

2. review the majority of valid notifications of hate speech in less than 24 hours.

The most recent evaluation campaign of the Code of Conduct [3] assessed in general a worse performance of the IT Companies especially in this last respect (only 64.4% of notifications were reviewed in less than 24 hours), confirming a troubling downward trend already observed in 2021.

Accurate, precise and efficient automated systems of hate speech detection are necessary for combating the phenomenon both with proactive action and in the process of reviewing user notification of occurrences of such violations.

Moreover, such systems should prove valuable for providing the data that is still needed to investigate the ecosystem of hate speech online and the magnitude of the phenomenon.

2 PROJECT GOAL

The goal of this project is to identify characteristics in the text which allow for the detection of Italian tweets containing hate speech (this category is meant to include expressions of racism, xenophobia and terrorist propaganda). So, we can define this task as a binary classification.

A secondary goal is to evaluate how insights obtained by our model can generalize across textual domains. In order to do that, we would like to test our model, trained exclusively on tweets, against Italian newspaper headlines.

Since tweets and newspapers' headlines exhibit very different stylistic and semantic characteristics (although they are both characterized by brevity), we expect overall a worse performance for the out-of-domain task. Nevertheless, we expect that the process will provide us with insights into the challenges associated with this transfer.

The project is based on the main task of the HaSpeeDe 2 shared task presented at EVALITA2020 [4].

3 AVAILABLE DATA

We will use the datasets made available[1] by the organizers of the Evalita 2020 shared task HaSpeeDe 2 [4].

The train set consists in 6839 Italian Tweets posted between October 2016 and May 2019 annotated for presence of hate speech. The test set includes both a corpus of 1263 Italian Tweets posted between January and May 2019 and a corpus of 500 Italian newspapers' headlines retrieved between October 2017 and February 2018 from the online editions of *La Stampa*, *La Repubblica*, *Il Giornale* and *Liberoquotidiano*. This second corpus will allow us to appraise how well the application generalizes across textual domains.

[1] https://github.com/msang/haspeede/tree/master/2020

Table 1: Distribution of Hate Speech labels.

	HS	Not HS	Total
Train	2766	4073	6839
Test Tweets	622	641	1263
Test News	181	319	500

Table 1 shows the distribution of hate speech labels in the training set and in each of the test sets.

The data sets are available in TSV format and contain three features for each record:

- `id`: numeric identifier for each document
- `text`: the body of the document
- `hs`: boolean value, whether the document contains HS (1) or not (0).
- `stereotype`: boolean value, whether the document contains a stereotype (1) or not (0) [might not be useful].

As part of the preprocessing performed by the organizers of the task, mentions and URLs recurring in the original documents were replaced with `@user` and `URL` placeholders.

We will also make use of the Italian Twitter embeddings [1] lexicon computed by ItaliaNLP Lab on a corpus of 50,000,000 tweets using the word2vec[2] toolkit. This lexicon consists of embeddings of 128 features for 1,188,949 tokens that were computed with the CBOW model with a symmetric context window of 5 tokens.

The embeddings were made available[3] by the ItaliaNLP Lab as a SQLite database containing a single table `store` with 130 columns:

- `key`, representing the token;
- one column for each dimension of the embedding;
- `ranking` storing the frequency rank of the token.

4 IMPLEMENTATION AND EVALUATION (TENTATIVE)

- Binary classification task (whether the document contains hate speech or not);
- Possible models [might change depending on what we'll cover on the course/what proves to be more interesting]: SVM (possibly testing different representations of the text as input: e.g. n-grams vs word2vec embeddings vs stylistic features), and fine-tuning pre-trained transformers such as DistilBERT [5]. The objective is to train multiple models using different text representations in order to enable a comparative analysis of two aspects: how the performance of a specific model is impacted by different representations, and the overall performance of different models.
- Evaluation metrics: Accuracy, Precision, Recall, F1-score.

For our evaluation we will make reference to two baselines proposed by the organizers of the shared task:

1. A classifier returning the most frequent class, which obtained a Macro-F1 0.3366 on the twitter test set and of 0.3894 on the news test set:
2. A Linear SVM using TF-IDF of unigrams and 2-5 char-grams, which obtained a Macro-F1 of 0.7212 for the twitter test set and of 0.621 for the news test set.

REFERENCES

[1] Andrea Cimino, Lorenzo De Mattei, and Felice Dell'Orletta. "Multi-task Learning in Deep Neural Networks at EVALITA 2018". In: *EVALITA Evaluation of NLP and Speech Tools for Italian*. Ed. by Tommaso Caselli et al. Torino: Accademia University Press, 2018, pp. 86–95. ISBN: 978-88-319-7842-2 978-88-319-7869-9. DOI: 10.4000/books.aaccademia.4527. URL: https://books.openedition.org/aaccademia/4527 (visited on 06/04/2024).

[2] European Commission. *The EU Code of conduct on countering illegal hate speech online*. URL: https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en (visited on 10/11/2024).

[2] http://code.google.com/p/word2vec/

[3] http://www.italianlp.it/download-italian-twitter-embeddings/

[3] Didier Reynders. *7th evaluation of the Code of Conduct*. Fact-sheet. European Commission, Nov. 2022. URL: https://commission.europa.eu/document/download/5dcc2a40-785d-43f0-b806-f065386395de_en?filename=Factsheet%20-%207th%20monitoring%20round%20of%20the%20Code%20of%20Conduct.pdf.

[4] Manuela Sanguinetti et al. "HaSpeeDe 2@ EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task". In: *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*. Ed. by Valerio Basile et al. Online: CEUR.org, 2020.

[5] Victor Sanh et al. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. 2020. arXiv: 1910.01108 [cs.CL]. URL: https://arxiv.org/abs/1910.01108.