Italian Hate Speech Detection

Miki Mizutani, Anna Fabris

Italian Hate Speech Detection problems

Complexity of task

- Tweets contain grammatical and spelling errors, slang
- Criteria for hate speech is ambiguous, even under human evaluation

Low resource language

- absence of large pretrained models
- fewer datasets

Dataset: talian Hate Speech Corpus

- 5156 tweets from Twitter users.
- Hate speech towards migrants and ethnic and religious minorities.
- Unbalanced with 15% hate speech.
- Partial preprocessing carried out.

Preprocessing

- Remove the reference to twitter users, links and hashtags.
- Normalise sequences of **at least 3** repeated characters with a maximum of two letters (e.g. hiiiiii → hii).
- Remove **numbers** and **punctuations**.
- Transform **emojis** into their aliases.
- Remove extra **white spaces** and any **left or right spacing**.

Models

Italian transformer

Italian BERT model.

Multilingual transformer

Roberta model trained on multilingual Twitter dataset.

English transformer with translated dataset

RoBERTa model trained on English tweets.

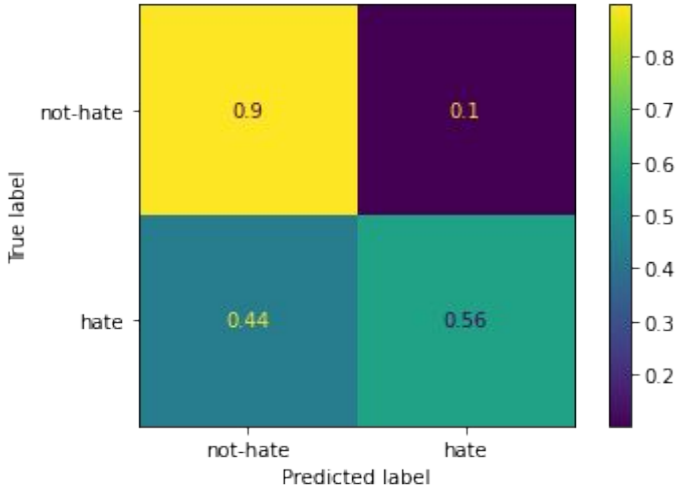
Hyperparameter Tuning

- Low learning rate 10^{-5}
- 128 batch size
- 7 epochs of training

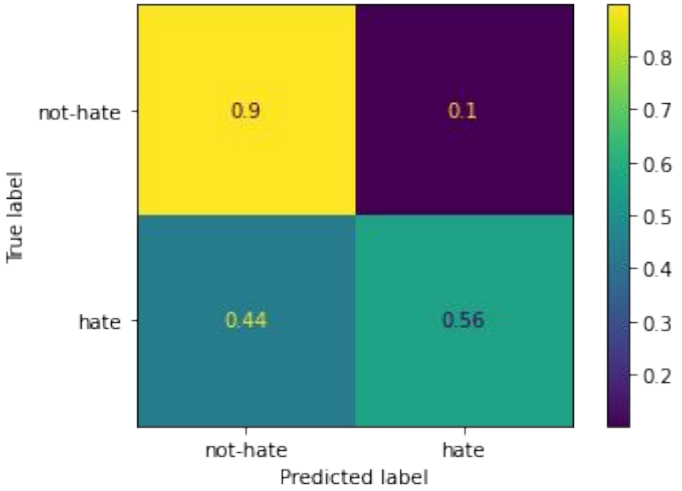
Results

Method	F1-score hate	F1-score not hate
Italian Transformer	31%	92%
Multilingual transformer	44%	92%
English transformer with translated dataset	28%	78%

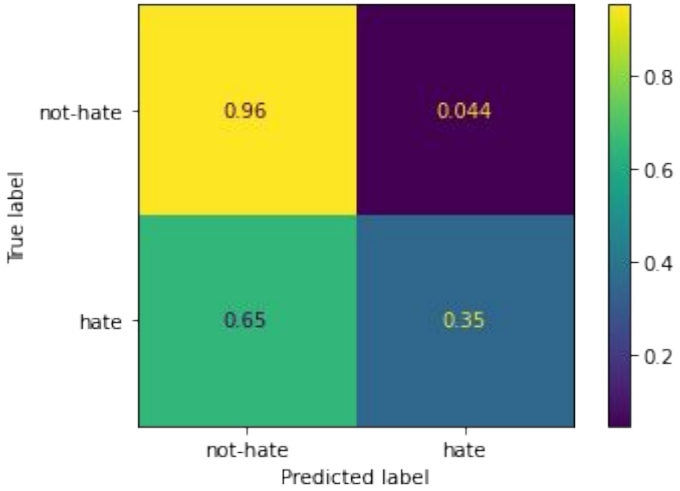
Confusion Matrix of Multilingual Transformer (Best Model)



Confusion Matrix of Multilingual Transformer with a weight bias of 1:2



No bias



Bias