Natural Language Processing Project

Anna Fabris
anna.fabris@studio.unibo.it

Miki Mizutani
miki.mizutani@studio.unibo.it

Academic Year: 2021 / 2022

Abstract

Hate speech detection of tweets as a classification problem has become a popular research topic in NLP. While hate speech online has become a worldwide problem, the focus has been primarily on detecting hate speech in English. With the absence of large, pre-trained models in non-English languages, and with the problem of ambiguity of what classifies as hate speech, training an effective, non-English hate speech model is difficult. In this project, we successfully train an Italian hate speech detecting transformer. We compare monolingual models in both Italian and English (where the Italian dataset is translated into English) and a multilingual model, and show that the Italian AlBERTo trained on Tweets in the Italian language had the best performance on this task.

Contents

1 Problem definition

In the age of social media, the ability to automatically detect hate speech has become a crucial task for social media giants such as Twitter. However, due to a variety of reasons, it is often a complex NLP task. Tweets often contain grammatical and spelling errors, slang, and lack context. Furthermore, the criteria for what constitutes as hate speech is ambiguous, even under human evaluation. The absence of large pretrained models and fewer datasets, further compounds the problem for non-English languages. Therefore, it is a challenging and important problem for NLP researchers to tackle this major issue.

For this project, we utilised the Italian Hate Speech Corpus[9] for a transformer based approach to detecting Italian hate speech on Twitter. We compare the Italian monolingual model, the multilingual model, and the English model with the translated dataset and show which technique can be useful technique for low-resource languages.

1.1 Related Works

Research into low-resource languages have seen a variety of monolingual and multilingual approaches. A multilingual approach was proposed by Aluru et al.[1] for multi-lingual embeddings in 9 different languages. In Italian in particular, Nozze et al.[6] presented a new hate speech dataset, along with a set of multi language models, showing that multilingual models outperform monolingual models trained solely on Italian datasets.

With regards to Italian datasets, aside from Sanguinetti et al.[9] which is a popular dataset for hate speech, Fersini et al.[3] and Bosco et al.[2] are also popular for hate speech related to misogyny.

Hate speech datasets have been criticised for their binary nature, such as by Vidgen et al.[10], who has built datasets with more descriptive labels than hate and non-hate. However, as the additional labels proposed by researchers are not normalized, it is difficult to join such datasets together.

2 Dataset

The dataset employed is the Italian Hate Speech Corpus[9]. It is a corpus of hate speech on Twitter towards migrants and ethnic and religious minorities (Roma and Muslims in particular).

The labels are highly unbalanced:

	Number of tweets	Percentage
Positive	811	15.7%
Negative	4345	84.3%
Total	5156	100%

Given this class unbalance, the results of the classifier would be skewed in favor of the minority class and hence potentially lead to misclassification which is one major problem of the dataset.

Every tweet was annotated for the categories of hate speech, aggressiveness, offensiveness, irony and stereotype, however for the purposes of this project, we only use hate speech as a label.

2.1 Preprocessing

Due to the short structure of tweets, users will typically compress semantic information using emojis or abbreviations. As such, it is necessary to pre-process the text to create a clearer semantic structure. Each tweet was pre-processed using the following methods:

- 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. heeeeeey → heey).
- Remove numbers.
- Remove punctuations.
- Transform emojis into their aliases.
- Remove extra white spaces.
- Remove any left or right spacing

The cleaned Italian dataset is then divided and shuffled into training, validation and test sets in the following way:

	Number of tweets	Percentage
Training set	3608	70%
Validation set	516	10%
Test set	1032	20%

3 Methods

We analyze different approaches to solve the aforementioned problems of lack of large pre-trained models and small number of datasets in the field of hate speech detection for non-English languages. First, we examine the monolingual transformer to detect hate speech in Italian. Second, we examine multilingual methods that use a larger English dataset to train a model for detecting hate speech in Italian texts using cross-language information transfer techniques. Thirdly, we examine a standard transformer for hate speech detection in English, applied to a dataset translated into English using Google Translate. And finally, we examine Italian monolingual AlBERTo trained on Tweets in the Italian language.

To model the data, 4 different transformer models were chosen: Italian Bert model, Twitter XLM Roberta Base, Bertweet Base Sentiment Analysis and Italian AlBERTo. Encoding was done with the pretrained encoders for each model.

For the transformer models, a low learning rate and a large batch size proved to be the most effective. We used a learning rate of 1e-5, a batch size of 128, with up to 7 epochs of training. The best model was evaluated using the F1 score, due to the unbalanced nature of the Italian dataset. Weight biasing was also tested with a 1:1.5 bias for non-hate and hate respectively, for the same reason. A confusion matrix was also plotted for each model to better visualise the predictions and errors.

3.0.1 Italian Monolingual model

For the monolingual Italian model we use a pre-trained cased Italian Bert model [4]. The source data for the Italian BERT model consists of a recent Wikipedia dump and various texts from the OPUS corpora collection.

3.0.2 Multilingual model

For the multilingual model, we used Twitter XLM Roberta Base [5], a Roberta model trained on multilingual Twitter dataset.

3.0.3 English model with translated dataset

For the English model we use Bertweet Sentiment Analysis [7] a RoBERTa model trained on English tweets.

The entire dataset has been translated using Google Translate document translation provided by Google Cloud Translation.

3.0.4 Italian model for Twitter languange understanding

We also tested a second Italian model an AlBERTo trained on Tweets in the Italian language [8].

4 Results

The test set is unbalanced as there are few hate examples. We thus show the performance of the models for both hate and not-hate classes.

	Precision		**Recall**		**F1 Score**	
	hate	not-hate	hate	not-hate	hate	not-hate
Italian Transformer	67%	87%	21%	98%	31%	92%
Multilingual transformer	58%	90%	35%	96%	44%	92%
English transformer with translated dataset	21%	88%	45%	70%	28%	78%
Italian transformer for Twitter understanding	62%	92%	55%	94%	58%	93%

Below are the confusion matrices of the models with the best hyperparameters.

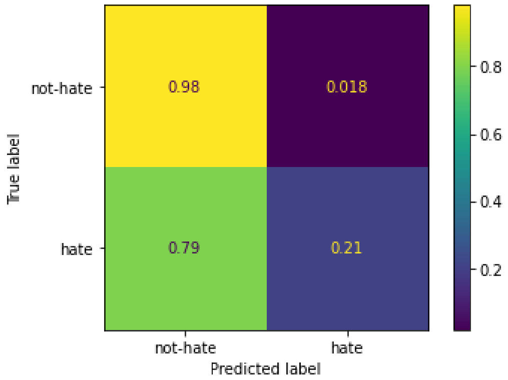

Figure 1: Confusion matrix of the Italian Transformer

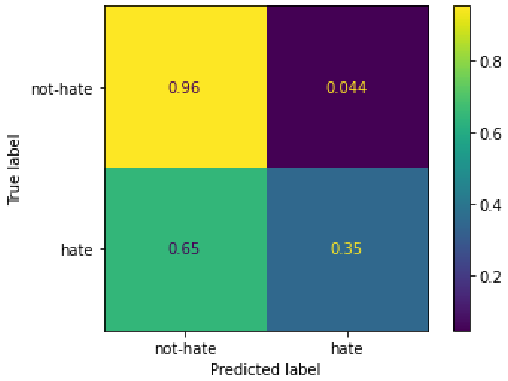

Figure 2: Confusion matrix of the multilingual Transformer

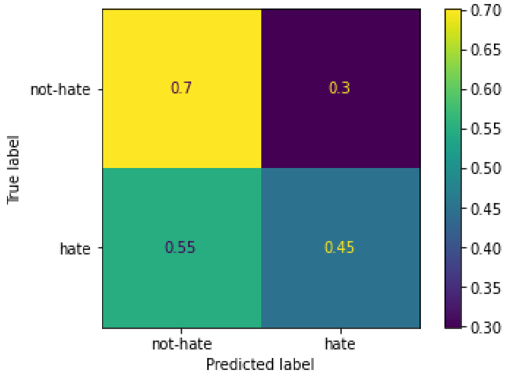

Figure 3: Confusion matrix of the English Transformer

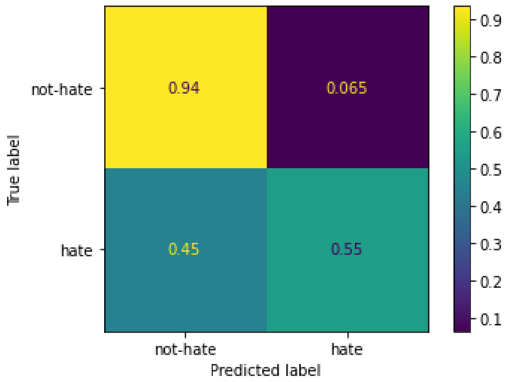

Figure 4: Confusion matrix of the Italian Transformer for Twitter language understanding

Based on our achieved results, we note certain interesting findings. Firstly, we note that the Italian model trained on Tweets has the the best overall results, second is the multilingual model. The first Italian model has comparable results, although the F1 score for the hate speech class is lower. The English model with the translated dataset has the worst results, which we hypothesise is due to the added error introduced through the translation. Overall, all models struggled with false positives in the hate class, which is to be expected in a highly unbalanced dataset.

To combat the unbalanced nature of the dataset, weight biasing was tested, however it gave mixed results. While applying a weight bias to the under-represented label (1:2 to no-hate, hate) gave better performance for the hate class, it inhibited the results of the no-hate class. Below is the confusion matrix for the Italian tweet transformer and multilingual transformer with a weight bias. While the number of false positives for the hate label goes down, the number of false positives for the non-hate label, goes up. In a real-world situation, we assume that precision in the hate speech class is more important, as the aim would be to remove all hate speech off social media platforms.

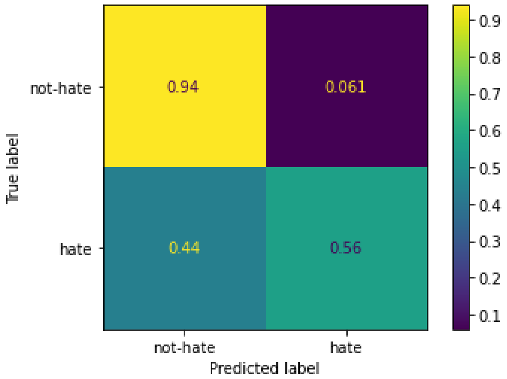

Figure 5: Confusion matrix of the Italian Twitter transformer with a weight bias of 1:2

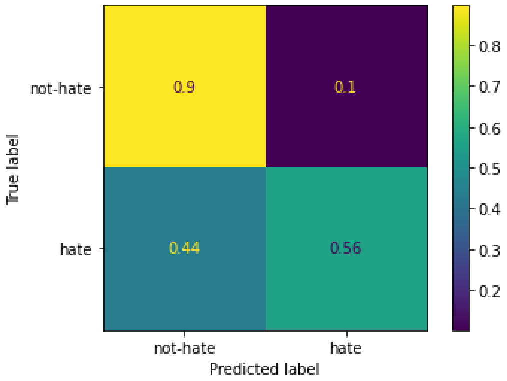

Figure 6: Confusion matrix of the Multilingual Transformer with a weight bias of 1:2

The Italian tweet model results are the best, but while there is a significant difference between weight bias and no weight bias in the multilanguage model, adding weight bias to the Italian tweet model results in few changes. This might be due to the model's already superior performance for the not-hate class.

Finally, we analyse our results with the Italian Tweet transformer with regards to the labels irony, aggressiveness, offensiveness and stereotype, that were provided in the dataset. While irony and stereotype had binary labels, aggressiveness and offensiveness were values on a scale of 0-1. For these labels, a threshold of 0.5 was applied to convert them to binary labels. Below are the confusion matrices for each label.

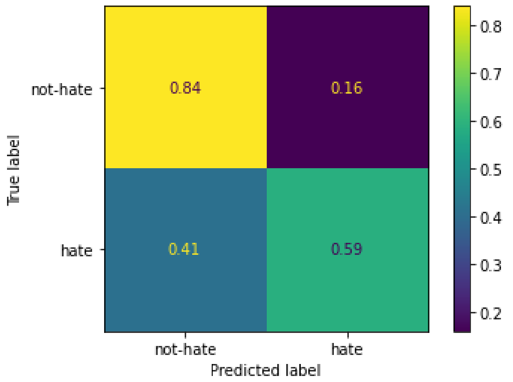

Figure 7: Confusion matrix of the irony label with the Italian Twitter transformer

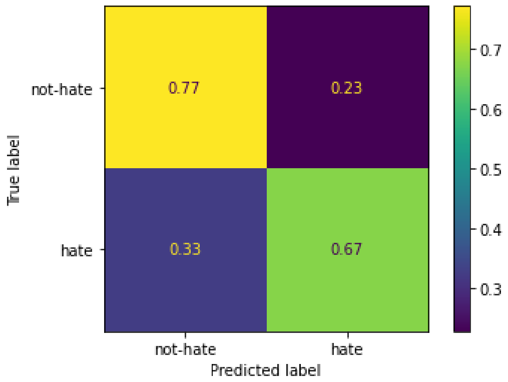

Figure 8: Confusion matrix of the aggressiveness label with the Italian Twitter transformer

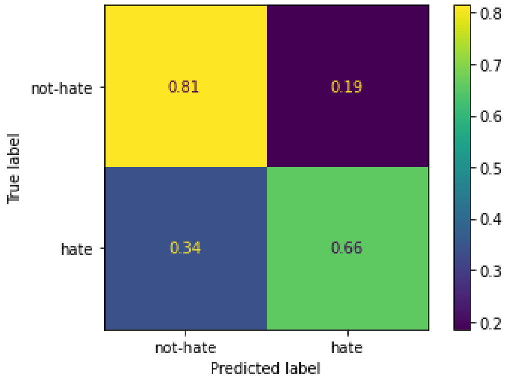

Figure 9: Confusion matrix of the offensiveness label with the Italian Twitter transformer

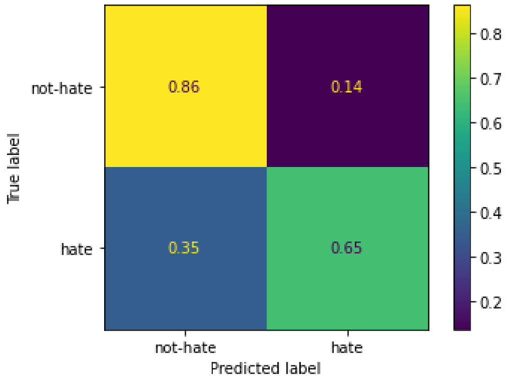

Figure 10: Confusion matrix of the stereotype label with the Italian Twitter transformer

Most significantly, the accuracy for the hate label is much lower for tweets that were labelled as ironic while the other labels had comparable results. We hypothesise that offensiveness, aggressiveness and stereotype, are easier to detect for models as they can be identified more lexically. On the other hand, irony in text can be difficult even for humans to identify, without context.

As future work for this problem, we believe that a dataset with additional information could be useful. Information such as demographics, location, timestamp, other tweets by the same user or the tweets before and after the offending tweet could all be helpful in increasing the understanding of the tweets.

Another possible improvement could be using a keyword-based approach. This is a fast and straightforward method for identifying hate speech. Text that contains potentially hostile keywords (like racial slurs) is found using database of derogatory terms. This method has many limitation: hateful slurs alone do not always qualify as hate speech, while hateful tweets without specific keywords wouldn't be identified. This usually results in a systems that with great precision but low recall. Given that all our transformers models have more problem identifying the hate class we predict that increasing using a keyword-based approach with the transformer could improve our results.

5 Conclusion

In this project, we compared an Italian monolingual model both trained and not trained on tweets, a multilingual model, and an English model for a binary hate speech classification task. We found that the Italian monolingual model trained on tweets gave the highest results, while the multilingual model was the second best. We showed that hate speech detection is a particularly difficult NLP task due to its ambiguity, especially with sentences that contain irony. We hope that this work helps research in future hate speech detection tasks in other low-resource languages, and to reach comparable results in these languages, as English.

References

[1] Sai Aluru et al. “A Deep Dive into Multilingual Hate Speech Classification”. In: Feb. 2021, pp. 423–439. ISBN: 978-3-030-67669-8. DOI: 10.1007/978-3-030-67670-4_26.

[2] Cristina Bosco et al. “Overview of the EVALITA 2018 Hate Speech Detection Task”. In: Jan. 2018, pp. 67–74. ISBN: 9788831978422. DOI: 10.4000/books.aaccademia.4503.

[3] Elisabetta Fersini, Debora Nozza, and Paolo Rosso. “Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI)”. In: *EVALITA@CLiC-it*. 2018.

[4] *Huggingface Italian Bert*. https://huggingface.co/dbmdz/bert-base-italian-cased. Accessed: 2022-08-20.

[5] *Huggingface Twitter XLM Roberta Base*. cardiffnlp/twitter-xlm-roberta-base. Accessed: 2022-08-20.

[6] Debora Nozza, Federico Bianchi, and Giuseppe Attanasio. “HATE-ITA: Hate Speech Detection in Italian Social Media Text”. In: *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*. Seattle, Washington (Hybrid): Association for Computational Linguistics, July 2022, pp. 252–260. URL: https://aclanthology.org/2022.woah-1.24.

[7] Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. *pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks*. 2021. arXiv: 2106.09462 [cs.CL].

[8] Marco Polignano et al. “AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets”. In: *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*. Vol. 2481. CEUR, 2019. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074851349&partnerID=40&md5=7abed946e06f76b3825ae5e294ffac14.

[9] Manuela Sanguinetti et al. “An Italian Twitter Corpus of Hate Speech against Immigrants”. In: *Proceedings of the 11th Conference on Language Resources and Evaluation (LREC2018), May 2018, Miyazaki, Japan*. 2018, pp. 2798–2895.

[10] Bertram Vidgen et al. “Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection”. In: Jan. 2021, pp. 1667–1682. DOI: 10.18653/v1/2021.acl-long.132.