LCTs at HODI: Homotransphobic Speech Detection on Italian Tweets

Davide Locatelli[1], Lorenzo Locatelli[2]

[1]Technical University of Catalonia, 31 Calle Jordi Girona, 08034 Barcelona, Spain
[2]University of Groningen, Broerstraat 5, 9712 CP Groningen, Netherlands

Abstract

Recent research highlighted the importance of employing language and culture-specific techniques to accurately detect homotransphobic speech. In this paper, we present our involvement in Subtask A of EVALITA 2023's HODI shared task [1], which specifically addresses the identification of homotransphobic content in Italian tweets. Our approach employs a classifier built upon pre-trained Italian word embeddings. Our approach achieves the best results in the shared task, and can serve as a valuable tool to combat this harmful phenomenon. We release our code at https://github.com/davidelct/hodi2023.

***Warning**: This paper contains examples of potentially offensive content. Profanities have been obfuscated with PrOf (https://github.com/dnozza/profanity-obfuscation) [2]*

Keywords

hate speech detection, homotransphobia, social media

1. Introduction

Social media platforms have revolutionized communication, providing a space for diverse viewpoints and opinions to be shared. While these platforms offer invaluable means of connection and expression, they have unfortunately also become breeding grounds for online harassment, particularly targeting minorities. This pervasive issue has raised significant concerns about the safety and well-being of the LGBTQIA+ community, which often faces homotransphobic harassment in digital spaces [3].

One of the challenges associated with combating online harassment is the ease with which users can freely express prejudiced views without immediate consequences. Compounding the problem, social media algorithms often contribute to the formation of echo chambers, where individuals are predominantly exposed to content that reinforces their existing beliefs [4]. Consequently, these algorithms can inadvertently perpetuate discriminatory attitudes and create an environment where hate speech thrives.

To address this pressing problem, the field of natural language processing (NLP) offers valuable resources that can effectively identify harmful online content and reduce its prevalence through automated hate speech detection systems. By leveraging NLP techniques, online moderators, who shoulder the responsibility of identifying and flagging dangerous content, can significantly alleviate the psychological strain associated with their role.

Recent research has shed light on the pervasive and complex nature of online homotransphobic hate speech [5, 6], showing a strong correlation between such hate speech and specific linguistic and cultural factors. This emphasizes the importance of targeted strategies that consider the linguistic context in which it occurs.

In this paper, we explore a promising approach to address the issue of homotransphobic hate speech on social media. Specifically, we leverage pre-trained word embeddings derived from large language models to build a classifier.

We use Subtask A of the HODI shared task [1] from the EVALITA 2023 workshop [7] to demonstrate that a classifier based on monolingual Italian word embeddings yields high results, highlighting how this approach can capture the nuances of the cultural factors at play. In Subtask A the goal is to predict whether a given tweet contains homotransphobic speech or not. We found that our approach achieves the highest results in the shared task.

The remainder of this paper is organized as follows. Section 2 describes the data used in this work, and the preprocessing techniques we employed. Our methodology, including the specifics of our experimental setup, is presented in Section 3. Section 4 showcases the results we obtained, while Section 5 contains a qualitative analysis of the errors made by the different models in our study. Section 6 concludes the paper discussing the implications of our research and proposing future directions to tackle homotransphobic hate speech on social media.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
davide.locatelli@upc.edu (D. Locatelli)
https://davidelct.com (D. Locatelli)
0009-0006-4194-4907 (D. Locatelli)

Table 1
Number of *homotransphobic* and *non-homotransphobic* tweets in the HODI Subtask A dataset.

Split	Homotransphobic	Non-Homotransphobic
Train	2,008	2,992
Test	511	489
Total	2,519	3,481

2. Data

Here, we present an overview of the data utilized in our study. This includes both the data released as part of the HODI challenge, as well as the data on which the models we utilized were pre-trained on. We did not undertake the pre-training step ourselves; nevertheless, we believe that describing the data is essential to offer a comprehensive understanding of the information to which the model has been exposed.

2.1. HODI Dataset

The HODI task organizers provided 6,000 Italian tweets manually labeled by expert annotators. For Subtask A, the annotators categorized tweets into two classes: homotransphobic or non-homotransphobic. The dataset was split into 5,000 tweets for training and 1,000 tweets for testing. To monitor the progress of our experiments, we reserved 200 tweets from the training set for validation.

The dataset statistics, as presented in Table 2.1, reflect a well-balanced distribution between the two classes across both the training and testing splits. This equilibrium enhances the reliability of our results and ensures that our model receives sufficient exposure to diverse instances of homotransphobic and non-homotransphobic language in Italian tweets during the fine-tuning process.

2.2. OSCAR Dataset

During the pre-training phase, the data utilized was the Italian corpus from the OSCAR dataset [8]. This particular collection of data is extensive, consisting of approximately 70GB of plain text. Specifically, it contains 210 million sentences and 11 billion words. The inclusion of such a vast amount of linguistic data ensures the model's exposure to a wide range of sentence structures, vocabulary, and syntactic patterns present in the Italian language.

3. Methodology

In this section we illustrate our approach, explaining both the data pre-processing steps we undertook, as well as the details of the models we utilized for Subtask A.

Table 2
Hyperparameters of models run1, run2 and run3

Hyperparameter	Value
N. epochs	{3, 5, 10}
Training batch size	16
Valid batch size	16
Warmup steps	500
Weight decay	0.01
Learning rate	2e-5

3.1. Data pre-processing

Our pre-processing consists of removing usernames, hashtags, and unnecessary white spaces from the tweets. To tokenize the text, we utilize the tokenizer associated with the pre-trained model that we describe in the next section.

3.2. Models

The three models used in our submission all consist of classifiers built on top of UmBERTo [9]. The three models all share the same hyperparameters (see Table 3.2), but they differ in the number of fine-tuning epochs on the HODI Subtask A data. Specifically:

Model run1 was fine-tuned for 3 epochs.

Model run2 was fine-tuned for 5 epochs.

Model run3 was fine-tuned for 10 epochs.

UmBERTo is a Roberta-base language model [10] pre-trained on Italian text using SentencePiece and Whole Word Masking techniques. For our classification tasks, we specifically utilized the `UmBERTo-Commoncrawl-Cased` version. [1] Using the HuggingFace Transformers library [11], we applied a classification head on top of the model outputs, which enabled us to fine-tune the base model on the HODI data for Subtask A.

The selection of the UmBERTo-Commoncrawl-Cased version offers enhanced compatibility with a wide array of text sources in comparison to alternative versions such as `Umberto-wikipedia-uncased-v1`. The latter model is pre-trained on a smaller dataset consisting mainly of Wikipedia posts, resulting in a narrower variety of text types compared to OSCAR. Furthermore, the version we selected retains the original casing of the text, which can provide significant insights especially in social media posts, where casing often serves as a means to convey

[1] Available at https://huggingface.co/Musixmatch/umberto-commoncrawl-cased-v1.

Table 3
Results of all model runs submitted to the HODI Subtask A. Our runs are underlined.

Team name	Model name	Macro F1 score
LCTs	run3	0.8108
LCTs	run2	0.8000
odang4hodi	run1	0.7959
DH-FBK	run1	0.7950
extremITA	run2	0.7942
odang4hodi	run2	0.7920
DH-FBK	run2	0.7837
odang4hodi	run3	0.7804
LCTs	run1	0.7709
extremITA	run1	0.7431
INGEOTEC	run1	0.7153
Team Tamil	run1	0.6735
baseline	run1	0.6691
SOVRAG	run3	0.6634
SOVRAG	run2	0.6334
SOVRAG	run1	0.6108
CHILab	run3	0.5528
CHILab	run1	0.5205
CHILab	run2	0.5199

strong emotions, opinions, and emphasis, and can prove as a valuable signal to detect hate speech.

To optimize our model, we employ the AdamW optimizer [12] and utilize a linear learning rate scheduler. In Table 3.2, we provide information about our experimental configuration, outlining the specific hyperparameters we selected.

4. Results

To assess the accuracy of our model's predictions, we employ the Macro F1 score as the evaluation metric. Table 4 reports the results of our three runs, as well as all other submissions to the HODI Subtask A.

We can observe that our approach is highly competitive in the shared task. Specifically, **Model run3** and **Model run2** achieve the highest and second-highest score in the competition, with over 0.80 Macro F1 performance. However, it should be noted that all models in the top five achieve over 0.79 Macro F1, and are within 0.2 point difference. While **Model run1** does not appear in the top five runs, it still achieves over 0.77 Macro F1.

Focusing only on our runs, it is evident that the performance improves as we extend the fine-tuning process, as demonstrated by the increment in the score with additional epochs. This observation highlights the positive impact of longer fine-tuning periods on the model's predictive capabilities. By allowing the model to undergo more epochs, we enable it to refine its predictions.

Table 4
Top words from the false negative examples across all three models.

Word	English translation	Count
F*mminiello	Effeminate gay man	20
F*mminielli	Effeminate gay men	13
Rotto	Broken	7
Culo	A*s	6
C*zzo	D*ck	5
Gay	Gay	5
C*lattone	F*dgepacker	5
R*cchione	F*ggot	4
Lesbiche	Lesbians	4
Tr*vioni	Tr*nsvestite	3

5. Error analysis

We divide the error analysis in two parts. First we consider examples that have been incorrectly categorized as not homotransphobic by all models, despite the gold label indicating the presence of homotransphobic speech. In other words, we consider false negatives across all models. This is so that we can gain an understanding of where our system would fail to protect LGBTQIA+ individuals online, highlighting directions for further refinements.

Then we analyze examples on which **Model run1** and **run2** failed to identify homotransphobia, but on which **run3** succeeded. This is to gain an understanding of the impact of extended fine-tuning.

5.1. False negatives

In total, 108 examples were false negatives across models, i.e. were wrongly classified as not homotransphobic by all three models. We report the top 10 words appearing in these examples in Table 5.1. It is interesting to note that the term "f*mminiello" and its plural form are the most frequently occurring words.

This observation is noteworthy as the word is primarily used in the Neapolitan dialect rather than being widely employed throughout Italy. It suggests that all models struggle with dialectal words that are infrequently encountered in its Italian pre-training corpus. Further investigation revealed that the fine-tuning data for HODI Subtask A only included two tweets containing such word, explaining why none of the models recognized this particular case.

The remaining words in the table consist of various slurs, such as "rotto in culo" (a combination of the third and fourth words), which translates to "assf*cked." This expression stigmatizes anal sex and, as it is predominantly used in its masculine form to insult men, it implies a negative connotation towards gay male sex. However, it is important to note that this expression is also commonly

Table 5
Top words from the tweets where model run3 improved compared to the other models.

Word	English translation	Count
Seduto	Sat down	4
F*mminielli	Effeminate gay men	3
Grandissimo	Very big	2
Figlio	Son	2
Casa	Home	2
F*mminiello	Effeminate gay man	2
GIOELE	First name (male)	2
MAGALDI	Last name	2
Problema	Problem	2
Verona	Verona (city)	2

used to insult non-gay individuals, making the identification of harassment towards LGBTQIA+ individuals more complex and context-dependent (e.g., considering the identity of the person being targeted). Nevertheless, it is worth mentioning that even when used to target non-LGBTQIA+ individuals, many people may still consider such expression to be homotransphobic, which is up for debate.

5.2. Improvements from extended fine-tuning

In total, 29 examples were correctly classified by **Model run3**, and incorrectly classified by the other two. We report the top 10 words appearing in these examples in Table 5.2.

We can observe that model run3 corrects a few of the false negatives containing the word "f*mminiello" described above, suggesting that more epochs allow the model to pick up on more subtle patterns present in the rest of the tweets.

Another interesting phenomenon is that of the words "GIOELE MAGALDI", which are a first and last name of an Italian male author, often insulted on social media with homotransphobic slurs. It is interesting to observe that model 3 was able to pick up on the harassment of an individual compared to the previous runs. This author is often insulted in all-caps tweets, which might have helped the model pick up on the aggressiveness of the language.

6. Conclusion

In this paper we described our approach to the HODI Subtask A [1] at EVALITA 2023 [7] on homotransphobic speech detection. The goal of our participation was to assess the effectiveness of using a simple classifier based on monolingual pre-trained word embeddings. We built our model on top of UmBERTo, an Italian version of BERT, pre-trained on a large amount of Italian data. We fine-tuned it using the HODI Subtask A data. We experimented by running the fine-tuning process for different number of epochs, and obtained high Macro F1 scores for all runs, around 0.8.

In future work, it would be worth comparing this performance with that of classifiers based on multilingual pre-trained word embeddings. Given the linguistic and culture-specific phenomena that characterize homotransphobic speech, it would be interesting to understand whether targeted monolingual embeddings yield better results than multilingual ones, potentially uncovering whether the former have a better time with nuanced edge cases.

While Italian is not a low-resource language, it would be also interesting to run this experiment with multilingual embeddings obtained from a dataset that does not include Italian, to understand whether the model can generalize from languages that exhibit similar phenomena as the target.

Acknowledgments

We thank the task organizers for setting up this shared challenge and providing the HODI dataset, a valuable resource for future work on this important area of research. Davide Locatelli is part of the INTERACT group of the Technical University of Catalonia, and is supported by the European Research Council under the European Union's Horizon 2020 research and innovation program (grant No. 853459). We gratefully acknowledge the computer resources at Artemisa, funded by the European Union ERDF and Comunitat Valenciana, and the technical support provided by the Instituto de Fisica Corpuscular, IFIC (CSIC-UV).

References

[1] D. Nozza, A. T. Cignarella, G. Damo, T. Caselli, V. Patti, HODI at EVALITA 2023: Overview of the Homotransphobia Detection in Italian Task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[2] D. Nozza, D. Hovy, The state of profanity obfuscation in natural language processing scientific publications, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, 2023.

[3] GLAAD, Social media safety index, 2022. URL: https://sites.google.com/glaad.org/smsi/platform-scores, accessed: 2023-07-22.

[4] M. Cinelli, G. D. F. Morales, A. Galeazzi, W. Quattrociocchi, M. Starnini, The echo chamber effect on social media, Proceedings of the National Academy of Sciences 118 (2021) e2023301118. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2023301118. doi:10.1073/pnas.2023301118.

[5] B. R. Chakravarthi, R. Priyadharshini, R. Ponnusamy, P. K. Kumaresan, K. Sampath, D. Thenmozhi, S. Thangasamy, R. Nallathambi, J. P. McCrae, Dataset for identification of homophobia and transophobia in multilingual youtube comments, ArXiv abs/2109.00227 (2021).

[6] D. Locatelli, G. Damo, D. Nozza, A cross-lingual study of homotransphobia on Twitter, in: Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 16–24. URL: https://aclanthology.org/2023.c3nlp-1.3.

[7] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[8] P. J. Ortiz Suárez, L. Romary, B. Sagot, A monolingual approach to contextualized word embeddings for mid-resource languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 1703–1714. URL: https://aclanthology.org/2020.acl-main.156. doi:10.18653/v1/2020.acl-main.156.

[9] L. Parisi, S. Francia, P. Magnani, Umberto: An italian language model trained with whole word masking, https://github.com/musixmatchresearch/umberto, 2020.

[10] L. Zhuang, L. Wayne, S. Ya, Z. Jun, A robustly optimized BERT pre-training approach with post-training, in: Proceedings of the 20th Chinese National Conference on Computational Linguistics, Chinese Information Processing Society of China, Huhhot, China, 2021, pp. 1218–1227. URL: https://aclanthology.org/2021.ccl-1.108.

[11] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.

[12] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.