CHILab at HODI: A minimalist approach

Irene Siragusa[1], Roberto Pirrone[1]

[1]Dipartimento di Ingegneria @ Università degli Studi di Palermo, Viale delle Scienze, Edificio 6 90129 - Palermo

Abstract

This technical report illustrates the system developed by the CHILab team for the competition HODI at EVALITA 2023. The key idea of the method we proposed for the HODI Subtask A - Homotransphobia detection, was to develop different systems arranged as suitable combinations of Pre-Trained Language Model (PTLM) for embedding extraction, neural architectures for further elaborations over the embeddings and a classifier. In particular dense layers, LSTM, BiLSTM and Transformers were used as neural architectures. The best performing system across the ones investigated in this report was made by embeddings extracted via AlBERTo coupled with a Transformer that reaches a macro-F1 score of 0.753.

Keywords

homotransphobia detection, transformer, language model

1. Warning

This paper contains examples of potentially offensive content.[1]

2. Introduction

The increasing interest for gender-inclusive and non-discriminatory language passes through its counterpart in hate speech that it is largely spreading in social networks, particularly against the LGBTQIA+ community. The NLP community is currently involved in developing systems for hate speech detection as in MAMI (Multimedia Automatic Misogyny Identification) [2] and EDOS (Explainable Detection of Online Sexism) [3] where the focus was on detection of misogyny and sexism, but these datasets are focused neither on Italian nor in detecting hate speech against people from the LGBTQIA+ community.

This paper introduces the architecture proposed by the CHILab team for the EVALITA 2023 campaign [4], and in particular as regards the Homotransphobia Detection in Italian task (HODI Subtask A - Homotransphobia detection) [5]. The general approach relies on encoding the text into suitable word embeddings that are processed via neural architectures like LSTM, BiLSTM or Transformers. Finally, the output classifier detects the presence of homotransphobic content.

We conceived our pipelines as “minimalist” architectures. No generative models [6, 7] where considered in this respect to derive embeddings. Moreover, we decided not to use fine-tuning in our PTLMs to stress the use of light networks to be trained with low computing resources. Finally, we set up a unique approach for all the tasks we have participated in EVALITA 2023.

The paper is arranged as follows: Section 2 reports a description of our systems along with data pre-processing, while results are reported and discussed in Section 3. Concluding remarks are in Section 4.

3. Description of the system

The data set by the HODI organizers contains 6000 Italian tweets annotated accordingly to the presence of homotransphobic content. Since the training set released for the competition was made up of 5000 samples, this was randomly split in a training and validation set, using a 80-20 ratio, resulting in 4000 and 1000 samples respectively.

3.1. Pre-processing

The `[URL]` tag, mention references, and retweet notes were removed since they were not considered meaningful: in particular, mentions are referred to anonymized accounts thus they add no special information. This was done after an analysis on the most cited words and hashtags[2]. As reported in Table 1, the `[URL]` tag is the most frequent one between classes and adds no information just like the anonymized mentions that are not reported. During this analysis it was interesting to notice that the most cited words are slurs directed to LGBTQIA+ members. Although a first idea for approaching the task was to look for slurs, the direct inspection of the data set shows clearly that slurs are not a good indicator of homotransphobic content. Slurs, in fact, are widely used

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT

irene.siragusa02@unipa.it (I. Siragusa); roberto.pirrone@unipa.it (R. Pirrone)

[1] Profanities have been obfuscated with PrOf (https://github.com/dnozza/profanity-obfuscation) [1]

[2] for this analysis all the words were reported in their lower case form

Table 1
Word distribution statistics over the dataset divided per label.

All tweets	freq	NH tweets	freq	H tweets	freq
culo	1283	culo	937	r*cchione	596
url	995	rotto	657	ch*cca	488
r*cchione	971	url	637	url	358
rotto	912	gay	544	culo	346
gay	700	c*zzo	398	rotto	255
ch*cca	688	r*cchione	375	gay	156
c*zzo	529	fare	328	isterica	151
fare	466	caghino	287	fare	138
solo	348	solo	231	c*zzo	131
me	310	me	211	quel	131

Table 2
Hashtags distribution statistics over the dataset divided per label.

All tweets	freq	NH tweets	freq	H tweets	freq
gay	15	pride	14	gioelemagaldi	9
pride	14	prelemi	13	conte	6
prelemi	14	eurovision	11	casalino	5
eurovision	13	gay	10	gay	5
tellonym	9	tellonym	8	tortura	5
gioelemagaldi	9	pridemonth	8	attacchi	4
draghi	8	dazn	7	draghi	4
pridemonth	8	meloni	6	biohacking	3
dazn	7	omofobia	5	fronte	3
conte	7	escita	5	intelligence	3

from the LGBTQIA+ people as a self-definition method suggesting a (re-)appropriation of the term itself [8], and obviously tweets of this kind cannot be considered as homotransphobic so the slur word loses its negative connotation, as in the tweet:

> firmato una fr*cia in sessione:([3]

here the term *fr*cia* does not have any negative connotation. Therefore any word-dependent consideration about the polarization of homotransphobic speeches cannot be made, as the presence of slur words does not convey negative content, i.e. slurs cannot be regarded as representative elements for separating classes. The same considerations hold for the hashtags as reported in Table 2 where the most frequent ones are neutral words. Hence, the hashtag symbol was removed and the subsequent word was kept along with its meaning inside the tweet.

Similar considerations were made for emojis: also in this case a strong polarization in the use of emojis is not found, in particular in the ones that are more associated with disgust and hate (Table 3). Since emojis are deeply used in social media communication, they were kept. Based on the statistics reported in Table 3, the emoticons contained in the data set where manually substituted with the corresponding most frequent emoji. As an example, the ":(" emoticon was translated in "😭" even if this is not the exact correspondence. This approach does not inject bias in the data set as the different emoticons were very few, while their rough meaning is preserved thus avoiding to consider them as mere sequences of punctuation marks. No further elaboration were made over the tweets: words were not reported to their lower case form, thus allowing a more accurate extraction of

Table 3
Emoji distribution statistics over the dataset divided per label.

All	freq	NH	freq	H	freq
😂	95	😂	45	😂	50
🤣	83	🤣	44	🤣	39
😭	23	😭	19	🤮	13
❤️	19	❤️	15	😅	9
👍	19	😍	13	👍	8
😅	17	👍	11	🌈	8
💩	16	🤔	9	❗	8
🤮	15	💩	9	😎	7
😍	14	😉	8	💩	7
😉	13	🥰	8	🤡	6

[3] signed by a queer during the exam session

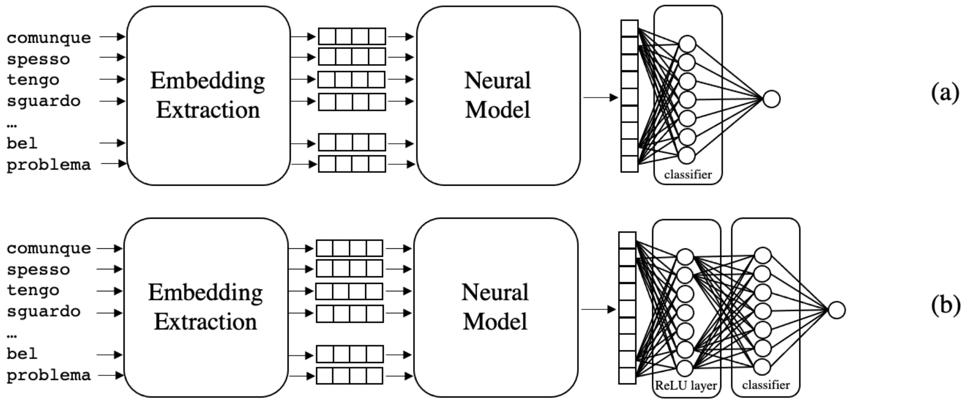

Figure 1: The first proposed architecture (a) has a module for embeddings extraction, a neural module for further processing on the extracted embeddings and a classifier. The second one (b) adds an additional ReLU dense layer.

embeddings for the case-sensitive PTLMs. As for emojis, uppercase texts has a specific meaning in social media communication in terms of prosodic and emotions interpretation [9, 10].

3.2. Network architectures

Different models were developed that share the same macro structure shown in Figure 1. The key idea was to stress, as much as possible, existent neural architectures for sequence processing, that are LSTM [11], BiLSTM and Transformers [12]. Those architecture are used to further process the extracted embeddings.

After pre-processing, the input sentences were padded to *maxLength* + 2 tokens where *maxLength* is the size of the longest sentence, and the remaining two tokens are respectively the `[CLS]` and the `[SEP]` one. Either a pre-trained language model or a static context-free embedding model were used for embedding extraction. In the last case, *fastText* [13] was used that generates a 300 tokens embedding, while a 768 tokens embedding is obtained as usual by the different PTLMs. We used the following Encoder-based Language Models in the experiments: BERT base multilingual cased [14], BERT base italian uncased [15], XLM-RoBERTa [16] and AlBERTo [17] provided by the HuggingFace Transformers library[4]. The embeddings were extracted from the last layer of the PTLMs without fine-tuning. Fine-tuning in these configuration is an option that is not taken into account since the main idea is to stress the use of light networks to be trained with low computing resources.

The extracted sequence of embeddings is fed into a neural module that consists of a LSTM or a BiLSTM or a Transformer[5]. The output feature vector has the same size of the word embedding with the exception of the BiLSTM that generates a double-length output. Finally, the feature vector is passed to a classifier made by either 300 or 768 linear units, depending on the length of the embedding, and a sigmoidal output to achieve binary classification (Figure 1.a). Some experimental configurations add an extra ReLU dense layer before the aforementioned classifier with exactly the same size. Those architectures are referred as LSTM-Deep, BiLSTM-Deep and Trasformer-Deep (Figure 1.b).

The illustrated architectures were trained only on the given data set using a machine equipped with two Intel Xeon E5 CPUs 96GB RAM and an NVIDIA TITAN Xp GPU 12GB RAM. Hyperparameters were selected as follows: dropout values in $\{0.1, 0.2\}$, batch size 32, Adam optimizer [18] with learning rate 0.01, and a Binary Cross Entropy loss. Models were trained for a maximum of 1000 epochs with a patience value of 50.

Different feature extractors were implemented using 1, 2 or 3 LSTM/BiLSTM/Transformer layers, but the best results were obtained by the single layer feature extraction modules. In addition the developed models are relative small, where the trainable parameters range from 1M to 10M.

4. Results

The best models during the evaluation window[6] were BERT-it/Transformer (run 1), AlBERTo/TransformerDeep (run 2) and AlBERTo/LSTM (run 3) and they

[4] https://huggingface.co/docs/transformers/index

[5] The corresponding architectures are named according the specific neural module

[6] they were the best models among the trained ones on our train/dev split

Table 4
The table collects the best obtained results with reference to the baseline value: the ones in italic place below the baseline; the underlined results are the ones generated removing stopwords from the data. XLM-RoBERTa, fastText and mBERT generate case-sentive embeddings.

	LSTM-Deep	**BiLSTM**	**Transformer**	**Transformer-Deep**
XLM RoBERTa	*0.565*	0.676	*0.650*	0.677
fastText	*0.554*	0.695	0.675	0.683
AlBERTo	0.717	0.745	0.705	**0.753**
mBERT	*0.333*	*0.630*	*0.338*	0.726
BERT-it	*0.642*	0.697	0.725	0.690

Table 5
The table collects the macro F1 results over the test set of the submitted models and their fixed versions (the starred ones). Result of the baseline model is also reported, along with the ranking and expected ranking position.

Run name	**Macro F1**	**Rank**
CHILab2*	0.753	10*
CHILab3*	0.745	11*
CHILab1*	0.725	13*
Baseline	0.669	13
CHILab3	0.553	17
CHILab1	0.521	18
CHILab2	0.520	19

placed ad the bottom of the rank and below the baseline [5]. Due to an internal error in the code of the training procedure, the submitted results are intrinsically wrong and for this reason, we repeated all the experiments using the correct architecture, after the release of the golden labels. An overview of all the developed models is reported in Table 4, while Table 5 shows the submitted runs, their fixed counterpart and the baseline value. In both tables the results refer to the F1-macro score over the test set and, although all possible configurations were run, in Table 4 we report the significant architecture, i.e. the configurations that placed above the baseline. The results show that the AlBERTo/Transformer architecture with a two dense layers classifier (Transformer-Deep) has the best performance, and it is expected to rank at the 10th place on the leaderboard.

Moreover, LSTM-Deep and BiLSTM models exhibit comparable performance: bi-directional sequence processing compensates for the reduced classifier's capacity. In general, the Transformer-Deep architectures performed better than the Transformer ones.

As it was expected, only the models based on *fastText* benefit from removing the stop words. The models using AlBERTo and BERT-it achieved almost the best results both in the training phase and in the evaluation, because the network can take advantage of PTLMs that are specifically fine-tuned on the target languages. In particular, AlBERTo was trained on a corpus of Italian tweets that share the same linguistic macro-structure of the data set proposed for HODI competition.

4.1. Error analysis

As suggested by the organizers of the shared tasks, an error analysis was performed particularly on the tweets that were mis-classified by the models reported in Table 4 that performs better with reference to the baseline. All classifiers agreed incorrectly on 40 tweets: the 80% of them were homotransphobic ones. Thanks to a direct analysis of their content, the following consideration can be made.

The very first consideration is that the majority of the mis-classified tweets contain slurs. As it has been shown in Section 3.1, slur words are widely used by the LGBTQIA+ people as self-reference without any discriminatory intent, so an automatic classifier may not recognize these shades of meaning as in:

> Fanculo Dolce & Gabbana non metto la roba fr*cia[7]

Moreover, many non-homotransphobic tweets share actually some linguistic similarities with the homotransphoic ones:

> DI ANORMALE c'è solo che una cripto ch*cca repressa e omofoba quale lei è #Pillon sia miserabile Senatore della Repubblica pagato dagli italiani e che peggio getta discredito sulla nostra Nazione con esternazioni quotidiane di puro, spregevole letame.[8]

Here some hateful content is reported towards a person that is considered homotransphobic. In those cases, the presence of hate speech is correctly detected but it does not meet the homotransphobic requirement.

[7] Fuck Dolce & Gabbana I do not wear f*g stuff

[8] ABNORMAL there is only that a repressed and homophobic crypto queer like you #Pillon is a miserable Senator of the Republic paid by the Italians and worse, discredits our nation with daily utterances of pure, despicable manure.

5. Conclusion

This paper reported the architectures developed by the CHILab team for HODI Subtask A promoted at the EVALITA 2023 campaign. Our models show that a relatively small classical pipeline made by embedding extraction plus further neural elaboration can have satisfactory performance in homotransphobic speech detection without the need of fine-tuning PTLMs, and using few computational resources. The use of such "minimalist" architecture is intended to allow for future development of compact explainable models where explicit linguistic knowledge is injected in the network to improve its performance.

Acknowledgments

This work is supported by the PO FESR 2014-2020 grant n. 086201000543, "SCuSi - Smart Culture in Sicily"

References

[1] 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.

[2] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, SemEval-2022 task 5: Multimedia automatic misogyny identification, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 533–549. URL: https://aclanthology.org/2022.semeval-1.74. doi:10.18653/v1/2022.semeval-1.74.

[3] H. R. Kirk, W. Yin, B. Vidgen, P. Röttger, SemEval-2023 Task 10: Explainable Detection of Online Sexism, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, 2023. URL: http://arxiv.org/abs/2303.04222. doi:10.48550/arXiv.2303.04222.

[4] 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.

[5] 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.

[6] G. Mialon, R. Dessì, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, B. Rozière, T. Schick, J. Dwivedi-Yu, A. Celikyilmaz, et al., Augmented language models: a survey, arXiv preprint arXiv:2302.07842 (2023).

[7] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, Z. Wu, D. Zhu, X. Li, N. Qiang, D. Shen, T. Liu, B. Ge, Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models, 2023. URL: http://arxiv.org/abs/2304.01852. doi:10.48550/arXiv.2304.01852, arXiv:2304.01852 [cs].

[8] E. Nossem, Queer, frocia, femminiellə, ricchione et al.–localizing "queer" in the italian context, Issue: gender/sexuality/italy, 6 (2019) (2019).

[9] M. Heath, Orthography in social media: Pragmatic and prosodic interpretations of caps lock, Proceedings of the Linguistic Society of America 3 (2018) 55–1–13. URL: https://journals.linguisticsociety.org/proceedings/index.php/PLSA/article/view/4350. doi:10.3765/plsa.v3i1.4350.

[10] S. Chan, A. Fyshe, Social and emotional correlates of capitalization on Twitter, in: Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, Association for Computational Linguistics, New Orleans, Louisiana, USA, 2018, pp. 10–15. URL: https://aclanthology.org/W18-1102. doi:10.18653/v1/W18-1102.

[11] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Computation 9 (1997) 1735–1780. URL: https://doi.org/10.1162/neco.1997.9.8.1735. doi:10.1162/neco.1997.9.8.1735. arXiv:https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. arXiv:1706.03762.

[13] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning word vectors for 157 languages, in: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[14] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[15] S. Schweter, Italian bert and electra models, 2020. URL: https://doi.org/10.5281/zenodo.4263142.

doi:10.5281/zenodo.4263142.

[16] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[17] M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, V. Basile, 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), volume 2481, CEUR, 2019. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074851349&partnerID=40&md5=7abed946e06f76b3825ae5e294ffac14.

[18] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).