A Multi-Task and Multilingual Model for Sexism Identification in Social Networks

Francisco Rodríguez-Sánchez, Jorge Carrillo-de-Albornoz, and Laura Plaza

UNED NLP & IR Group, Calle Juan del Rosal, 16. 28040 Madrid, Spain
frodriguez.sanchez@invi.uned.es,{jcalbornoz,lplaza}@lsi.uned.es

Abstract. Exposure to sexist content has serious consequences for women's life and limits their freedom of speech. In this paper, we present a multilingual system based on pre-trained transformers and compare single-task to multi-task learning to identify sexism in social networks. Our methods have been evaluated in the framework of our participation in the EXIST shared task at IberLEF 2021 [1] obtaining promising results despite sharing parameters for both languages and tasks.

Keywords: Sexism detection · NLP · Transformers · Multi-task learning.

1 Introduction

The development of web technologies has enabled the interaction between people from many different countries and backgrounds. With more than 4 billion people around the world now using social media each month [2], social networks are undoubtedly one of the most important ways of communicating. Although we can not deny the positive effects of this global communication, anonymity and accessibility have made the expression of discriminatory and sexist discourses easy and unpunished. In this context, inequality and discrimination against women that remain embedded in society are increasingly being replicated and spread online.

The Oxford English Dictionary defines sexism as "prejudice, stereotyping or discrimination, typically against women, on the basis of sex". Therefore, sexism is expressed in very different forms that do not always express hostility or hate. Subtle forms of sexism can be as pernicious as other forms of sexism and affect women in many facets of their lives. According to [3], non-hateful sexism can affect women's psychological well-being by decreasing their comfort, increasing their feelings of anger and depression, and decreasing their stated self-esteem. Similarly, [4] found a relationship between the experience of non-violent sexism and posttraumatic stress disorder.

IberLEF 2021, September 2021, Málaga, Spain.

Detecting sexist content is still a difficult task for social media platforms. For instance, Amnesty International published a report [5] where they describe Twitter as a “toxic place” for women. According to this report, Twitter is promoting violence and hate against people based on their gender. The report also suggests that Twitter is failing to protect women against harassment and it could harm their freedom of speech. Recently, members of the U.S. Congress asked Facebook to do more to protect women in their platform. According to some lawmakers, social media has become “the number one place” in which psychological violence is perpetrated against female parliamentarians [6]. The seriousness of the problem, combined with the quick spread of online information, especially in social networks, has made these harassment behaviours extremely dangerous so that solutions are required to perform a faster and even better user generated-content moderation, or to serve as a tool that helps human moderators to reduce the volume of sexist content still present in online platforms.

In this paper, we describe our participation in the EXIST task at IberLEF 2021, a sexist language detection task in two different languages. The challenge was articulated in two different tasks: task 1 is a binary classification to determine whether a text is sexist or not, while task 2 is a finer-grained classification devoted to distinguishing different subtypes of sexism. We propose a multilingual system based on pre-trained transformers and experiment with single-task and multi-task approaches to jointly address the task of sexist language detection. We take advantage of the fact that both tasks are semantically connected to test whether both tasks can be simultaneously learned and one task can benefit the other using a multi-task framework. To the best of our knowledge, no previous work has employed this technique to identify sexism in social networks. Our single-model approach achieved competitive results, with a performance close to top-performing systems despite sharing parameters for both languages and tasks.

The rest of this paper is organized as follows: in section 2, we discuss related works. In section 3, we describe the classification system. Results and analysis are presented in section 4. Finally, the conclusions and future works are given in section 5.

2 Related work

The detection of hate speech and misogyny are tasks that are closely connected and often confused with sexism detection [7]. Substantial work has been devoted to the detection of hate speech in recent years but few works have faced sexism detection. Most of them have dealt with sexism as the detection of hate speech against women or misogyny [8]. Consequently, they have worked with hostile and explicit sexism, overlooking subtle or implicit expressions of sexism. An exception is the approach proposed by [7], where authors released the first Spanish corpus of sexist expressions in Twitter, the MeTwo dataset. They also compared Machine Learning (ML) methods to detect sexism and discussed the generalization of their approach with respect to misogyny detection systems.

Recently, the IberEval competition focused on the automatic identification of misogyny in Twitter [8]. Teams were proposed to identify misogynist tweets both in Spanish and English. Approaches presented to the competition were mainly based on supervised machine learning on different textual features (such as unigrams and bigrams, sentiment-based information, or syntactic categories) or user-based features (such as the number of retweets, followers, etc.) [9–11]. The use of lexical resources for extracting signals (such as swear word count, or sexist slurs presence) showed excellent performance in the task [12]. Deep learning methods were explored only by one team along with word embedding features [13].

The appearance of multilingual transformers has shifted the trend in natural language processing, with many positive experimental results for hate speech detection. For instance, [14] explored the feasibility of detecting misogyny in three different languages using the multilingual Bidirectional Encoder Representation from Transformer (multilingual BERT or mBERT) [15]. Another example is found in [16], where authors presented an ensemble model of individual transformers as the winner solution in the shared task "Offensive Language Identification in Dravidian Languages" at EACL 2021.

Multi-task learning (MTL) has proven successful in many Natural Language Processing (NLP) problems, as illustrated in the overview of [17]. In this paradigm, multiple tasks are simultaneously learned by a shared model offering advantages like improved data efficiency, reduced overfitting through shared representations, and fast learning by leveraging auxiliary information. Only a few studies are using MTL to detect hate speech language. [18] employed emotion detection as the auxiliary task to address the detection of abusive language. Another example can be found in [19], where a MTL approach was applied to detect hostile content.

Although multilingual and multi-task models have been tested as end-to-end solutions for several tasks related to hate speech, to the best of our knowledge, no previous work has explicitly used these techniques for the sexism detection task.

3 EXIST 2021: sEXism Identification in Social neTworks

The shared task EXIST 2021 at IberLEF 2021 [20] asked participants to classify "tweets" and "gabs" in two different languages, English and Spanish. The objective of the shared task is to develop methodologies and classification systems to detect sexist messages according to the following two tasks:

- Task 1: It is a binary classification task, where every system should determine whether a text or message is sexist or non-sexist.
- Task 2: Once a message has been classified as sexist, the second task aims to categorize the message according to 5 types of sexism: Ideological and inequality, Misogyny and non-sexual-violence, Objectification, Sexual violence, Stereotyping and dominance.

Task 1 is evaluated in terms of accuracy, while for Task 2 the evaluation consists in the macro-average of the F1-scores on the 6 classes: Non-sexist, Ideological and inequality, Misogyny and non-sexual-violence, Objectification, Sexual violence, Stereotyping and dominance. Each participating team could submit a maximum of 6 runs, 3 runs for each task.

Two different datasets were shared during the challenge. In total, the organizers provided 6977 tweets for training and 4368 texts for testing composed by 3.386 tweets and 982 gabs. The organizers ensured class balancing according to task 1, while the distribution of data for task 2 was relatively unbalanced, reflecting a more natural distribution of sexist content.

4 System description

In recent years, transformer-based language models like BERT [15] and its variant RoBERTa [21], have become the state of the art for most NLP tasks. In particular, multilingual versions of these systems have shown surprising cross-lingual capabilities, even among languages that do not share scripts [22].

For our work, we fine-tuned three different state-of-the-art multilingual transformer models: mBERT, XLM-RoBERTa [23], and XLM-Twitter [24]. mBERT shares the same training as single-language BERT but using a concatenated dataset of 104 languages, XLM-RoBERTa (XLM-R) was trained on data from 100 and XLM-Twitter (XML-T) makes start from XLM-R and continue pre-training on a large corpus of Twitter in 30 languages.

While, in most cases, the multilingual models are trained and tested independently for each language and do not combine different languages in a single evaluation, our approach allows us to tackle the task for both languages at the same time and share the same model.

4.1 Single-task model

As the tasks are evaluated independently, we have explored transformer models for each task independently and will be referring to them as single-task models. Figure 1 shows the model architecture for this approach. On top of the transformer model, we added a linear layer to minimize loss function in our particular task. In particular, we used cross-entropy loss for both tasks, with two and six labels respectively: a binary problem for task1 and a multi-class classification with 5 types of sexism and non-sexist for task 2.

4.2 Multi-task learning with learnable parameter

To exploit the fact that both tasks share the same data distribution and are semantically connected, we propose to learn a model jointly on both of them. Figure 1 shows the model architecture for this approach. Specifically, we consider hard parameter sharing [17] among both tasks using a base model, followed by

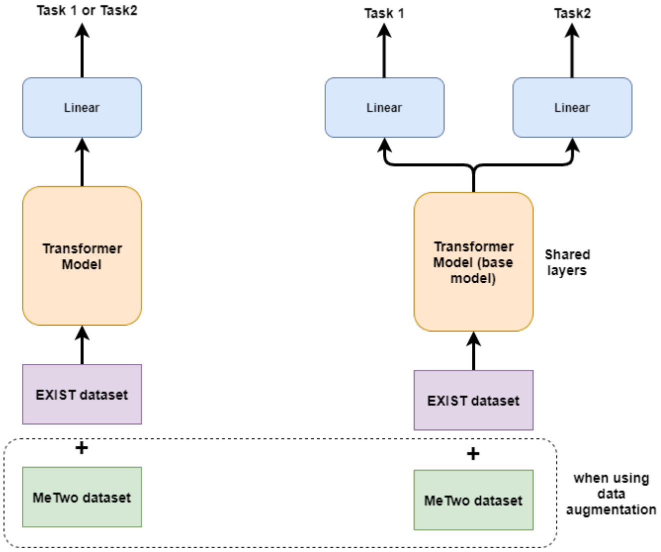


Fig. 1. Left: Single-task model. Rigth: Task 2 Multi-task model

two linear layers for each classification task. As base model, we employed all transformer models previously described in this section (section 4).

We experimented with the inclusion of a learnable parameter α to control the importance we place on each task in the multi-task learning framework. In particular, we compute loss with the following expression:

$$L = \alpha L_{TASK1} + (1 - \alpha) L_{TASK2}$$

Where L_{TASK1} and L_{TASK2} are cross-entropy losses for each task. Since in our problem both tasks are equally important, we set an initial value of $\alpha = 0.5$.

4.3 Data augmentation with MeTwo dataset

Data Augmentation is a quite popular solution to improve systems generalization by generating slight variants of the given dataset and is extremely useful for small datasets [25].

For our approach, we did some experiments concatenating the MeTwo dataset [7] to the EXIST dataset. In particular, we removed all tweets from the "DOUBTFUL" class in MeTwo and used the "SEXIST" and "NON-SEXIST" labels to perform this experiment for task 1. For multi-task experiments, tweets from MeTwo did not contribute to task 2 loss. Finally, since MeTwo is considerably unbalanced to the "NON-SEXIST" label, we balanced both classes.

5 Results and analysis

5.1 Experimental setup and preprocessing

All the experiments were performed using Pytorch [26] and HuggingFace [27] Transformers library. As the implementation environment, we used a NVIDIA Tesla T4 GPU. Optimization was done using Adam [28] optimizer with an initial learning rate of 2^{-5} and a linear weight decay of 0.01 for training single-task and multi-task models. We trained all models with a batch size of 16 for 20 epochs with an early stopping of 8 epochs. We make our code publicly available at Github [29].

The only preprocessing step before feeding the input to the transformer tokenizers was converting to lowercase, replacing mentions, hashtags, and URLs with a keyword, and removing punctuation signs.

To evaluate our systems, we trained all models on 70% of the training data, and held out the remaining 30% for validation.

5.2 Results

Here, we report the performance of the approaches described in the previous section. Table 1 summarizes the results obtained for different experiments in the validation set, they are reported in terms of accuracy and macro-f1. We observe that, among the individual transformer models, the best performance is obtained using XLM-T. It can be due to the fact that it is pre-trained using data from Twitter, the same datasource of our task.

In the case of multi-tasking approaches, classifiers perform well, having small differences with respect to single-task systems for task 1 and outperforming them for task 2. Regarding data augmentation using the MeTwo corpus, we can observe that it generally improves results for task 2, which could suggest that adding instances from task 1 improves results in task 2. It also should be noted that the inclusion of a learnable parameter α to control the importance we place on each task slightly improves results for task2.

We presented our three classifiers to the challenge using different approaches so that we could compare their performance in the test set. In particular, we sent a single-task classifier and two multi-task systems, using data augmentation and a parameter to control the importance of the tasks.

Table 2 illustrates the results obtained in the competition, where the results are reported in terms of accuracy for task 1 and macro-f1 for task 2. Regarding task 1, our single-task multilingual classifier performs quite well, achieving performances comparable to the winning teams. Similarly, our multi-task model performs fairly well, having a difference of around 2% with respect to the best result.

For task 2, most participants achieved relatively low results, showing the difficulty of this task. The multi-task approach yielded our best results and it stays in the top cluster of the competition (11 out of 63 runs). Unlike our experiments, using data augmentation did not perform well in the test set for

Table 1. Experimental results in the validation set

	Task 1		Task 2	
	Accuracy	Macro-F1	Accuracy	Macro-F1
mBERT-single-task	0,749	0,749	0,624	0,541
mBERT-multi-task	0,745	0,744	0,638	0,536
mBERT-multi-task-and-metwo-balanced	0,74	0,74	0,63	0,526
XLM-R-single-task	0,734	0,73	0,64	0,55
XLM-R-multi-task	0,773	0,773	0,649	0,535
XLM-R-multi-task-and-metwo-balanced	0,773	0,773	0,655	0,55
XLM-R-multi-task-learnable-parameter (run 3)	0,773	0,773	0,65	**0,575**
XLM-T-single-task (run 1)	**0,786**	**0,785**	0,662	0,572
XLM-T-multi-task	0,779	0,779	0,65	0,57
XLM-T-multi-task-and-metwo-balanced (run 2)	0,759	0,757	**0,667**	**0,575**
XLM-T-learnable-parameter	0,779	0,777	0,666	0,572

task 2. This could be due to the inclusion of Gab in the test set, which is biased towards aggressive sexism.

Once the evaluation phase was over, organizers shared the labels for the test set in case participants wanted to perform further tests. We added two extra experiments to table 2 using two models we did not present to the competition. As we can see, we would have obtained slightly better results for task 2. As we observed in our experiments, multi-task approaches yield better results for task 2 than single-task models.

Table 2. Official results EXIST test set

	Task 1		Task 2	
	Accuracy	Run Rank	Macro-F1	Run Rank
Rank-1	***0,784***	1	***0,5787***	1
Majority Class (baseline)	0,6845	66	0,4778	62
SVM TFIDF (baseline)	0,522	52	0,522	51
XLM-T-single-task (run 1)	**0,772**	**7**	0,544	15
XLM-T-multi-task-and-metwo-balanced (run 2)	0,7324	29	0,5246	22
XLM-R-multi-task-learnable-parameter (run 3)	0,7571	17	**0,5509**	**11**
XLM-T-multi-task	0,764	-	0,554	-
XLM-T-multi-task-learnable-parameter-concat-metwo	0,747	-	0,553	-

5.3 Error analysis

Although we achieve interesting results, all models are still making some mistakes. To understand better the source of the failures, we have performed a deep analysis of model errors. In particular, we further investigate the results of the single-task XLM-T model for each task.

Figure 2 displays the confusion matrix for tasks 1 and 2. Regarding task 1, the non-sexist class performs worse than the sexist one. For task 2, most errors

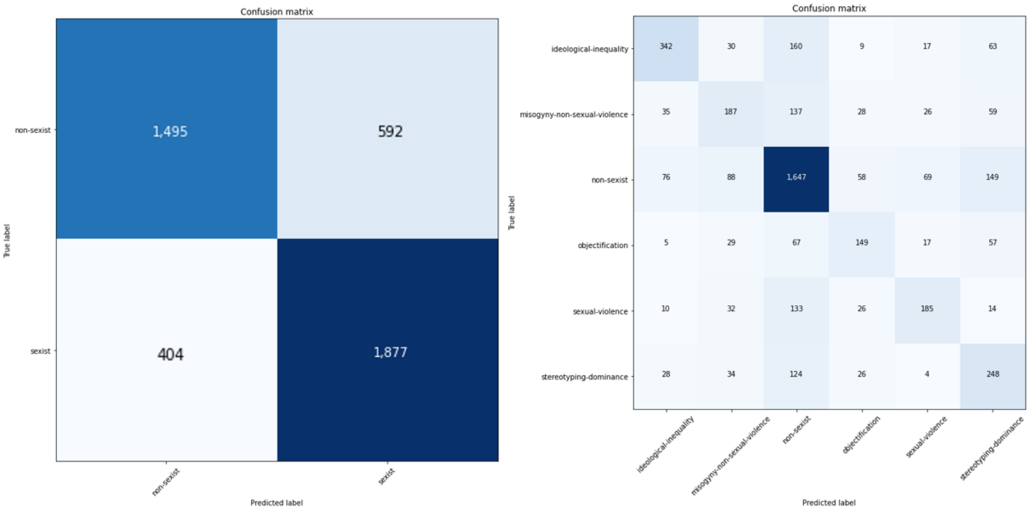
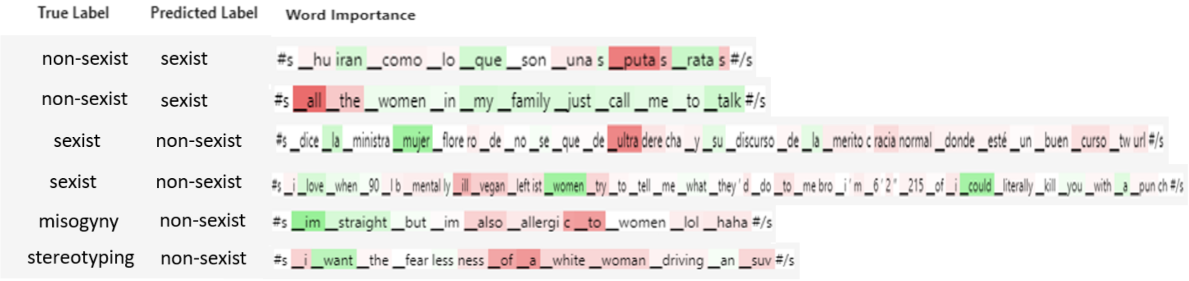


Fig. 2. Left: Task 1 confusion matrix. Rigth: Task 2 confusion matrix



True Label	Predicted Label	Word Importance
non-sexist	sexist	#s _hu iran _como _lo _que _son _una s _puta s _rata s #/s
non-sexist	sexist	#s _all _the _women _in _my _family _just _call _me _to _talk #/s
sexist	non-sexist	#s _dice _la _ministra _mujer _flore ro _de _no _se _que _de _ultra dere cha _y _su _discurso _de _la _merito c racia normal _donde _esté _un _buen _curso _tw url #/s
sexist	non-sexist	#s _i _love _when _90 _l b _mental ly _ill _vegan _left ist _women _try _to _tell _me _what _they ' d _do _to _me bro _i ' m _6 ' 2 " _215 _of _i _could _literally _kill _you _with _a _pun ch #/s
misogyny	non-sexist	#s _im _straight _but _im _also _allergi c _to _women _lol _haha #/s
stereotyping	non-sexist	#s _i _want _the _fear less ness _of _a _white _woman _driving _an _suv #/s

Fig. 3. Word importance

come from the misogyny-non-sexual-violence and stereotyping-dominance. We attribute this to the heterogeneity of these classes thus many types of sexist attitudes could be part of them. For instance, the sentence “Some woman are so toxic they don’t even know they are draining everyone around them in poison. If you lack self awareness you won’t even notice how toxic you really are” and “They refuse to arrest the separatist who has broken the nose of a woman for removing ties” are both instances of the misogyny-non-sexual-violence class, but for different reasons. On the contrary, ideological-inequality is a more homogeneous group and the performance is better.

To analyze the reasons behind the errors of our model, we used the library transformers-interpret [30] to have more information about the importance of each token towards the predicted class. Figure 3 shows the word importance for some errors examples. In this figure, red means that the token is pushing towards the “incorrect” (and predicted) class, whereas green pulls towards the correct class. As we can see, it turns out that numerous spurious correlations are learned by our classifier: words such as “puta” or “all” trigger the sentence as sexist. Similarly, the existence of words such as “ill” or “vegan” pushes the prediction of the 4th sentence towards non-sexist. The last two examples are related to errors for task 2. Both are cases where the classifiers fail to detect the type of sexism because of the appearance of irrelevant terms like “straight” and “want”.

6 Conclusions

In this paper, we have described a classification model for sexist language detection in a multilingual scenario. We also compared single-task to multi-task approaches and experimented with data augmentation techniques using a corpus from the same domain. The results obtained in the framework of the EXIST 2021 competition are promising since our single-model approach had close performance to top-performing systems despite sharing parameters for both languages and tasks. Furthermore, the results show how our model fitted spurious correlations for certain terms that must be carefully analyzed with more experiments.

As future work, we plan to experiment with the inclusion of affective lexicons to improve the automatic detection of sexism. It is also important to note that the strategy to construct the dataset is keyword-based, which can introduce natural biases towards certain sexist terms. Thus, bias mitigation techniques could be useful to improve performance.

Acknowledgments

This work was supported by the Spanish Ministry of Science and Innovation under Project Misinformation and Miscommunication in Social Media (PGC2018-096212-B-C32).

References

1. Montes, M., Rosso, P., Gonzalo, J., Aragón, E., Agerri, R., Álvarez-Carmona, M.Á., Álvarez Mellado, E., Carrillo-de Albornoz, J., Chiruzzo, L., Freitas, L., Gómez Adorno, H., Gutiérrez, Y., Jiménez Zafra, S.M., Lima, S., Plaza-de-Arco, F.M., Taulé, M.: Proceedings of the iberian languages evaluation forum (iberlef 2021). In: CEUR workshop (2021)
2. Datareportal report, https://datareportal.com/reports/digital-2020-october-global-statshot. Last accessed 25 May 2021
3. Swim, J.,Hyers, L., Cohen, L., Ferguson, J.: Everyday Sexism: Evidence for Its Incidence, Nature, and Psychological Impact From Three Daily Diary Studies. Journal of Social Issues **57**(1), 31–53 (2001)
4. Berg, H.: Everyday Sexism and Posttraumatic Stress Disorder in Women. Violence Against Women **12**(10), 970–988 (2006)
5. Amnesty International report, https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1/. Last accessed 25 May 2021
6. Reuters article, https://www.reuters.com/article/us-facebook-women-politics-idUSKCN2522KK. Last accessed 25 May 2021
7. Rodríguez-Sánchez, F., Carrillo-de-Albornoz, J, Plaza, L.: Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data. IEEE Access **8**, 219563–219576 (2020)
8. Fersini, E., Anzovino, M., Rosso, P.: Overview of the Task on Automatic Misogyny Identification at IberEval. In: Proceedings of 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with SEPLN 2018), pp.57-–64. CEUR-WS.org (2018)
9. Canós, J., S., Misogyny identification through SVM at IberEval 2018. In: Proceedings of Human Language Technologies for Iberian Languages (IberEval 2018), pp. 229—233. (2018)
10. Nina-Alcocer, V.: AMI at IberEval2018 automatic misogyny identification in Spanish and English tweets. In: Proceedings of Human Language Technologies for Iberian Languages (IberEval 2018), pp. 274—279. (2018)
11. Frenda, S., Ghanem, B.: Exploration of misogyny in Spanish and English tweets. In: Proceedings of Human Language Technologies for Iberian Languages (IberEval 2018), pp. 260—267. (2018)
12. Pamungkas, E., W.: Exploiting lexical knowledge for detecting misogyny in English and Spanish tweets. In: Proceedings of Human Language Technologies for Iberian Languages (IberEval 2018), pp. 234-–241. (2018)
13. Goenaga, I., Atutxa, A., Gojenola, K.,Casillas, A., Ilarraza A., Ezeiza, N., Oronoz, M., Pérez, A., Perez-de-Viñaspre, O.: Automatic misogyny identification using neural networks. In: Proceedings of Human Language Technologies for Iberian Languages (IberEval 2018), pp. 249-–254. (2018)
14. Pamungkas, E., W., Basile, V., Patti, V.: Misogyny Detection in Twitter: a Multilingual and Cross-Domain Study **57**(6), (2020)
15. Devlin, J., Chang, M. W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019), pp. 4171—4186. (2019)
16. Saha, D., Paharia, N., Chakraborty, D., Saha, P., Mukherjee, A.: Hate-Alert@DravidianLangTech-EACL2021: Ensembling strategies for Transformer-based Offensive language Detection. In: Proceedings of the First Workshop on

Speech and Language Technologies for Dravidian Languages (EACL 2021), pp. 270-–276. (2021)

17. Crawshaw, M.: Multi-Task Learning with Deep Neural Networks: A Survey, arXiv (2020)
18. Rajamanickam, S., Mishra, P., Yannakoudakis, H., Shutova, E.: Joint Modelling of Emotion and Abusive Language Detection. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), pp. 4270—4279. (2020)
19. Kamal O., Kumar, A., Vaidhya, T.: Hostility Detection in Hindi Leveraging Pretrained Language Models. In: First International Workshop, Combating Online Hostile Posts in Regional Languages during Emergency Situation 2021, Collocated with AAAI 2021, (CONSTRAINT 2021), pp. 213–223. (2021)
20. Rodríguez-Sánchez, F., Carrillo-de-Albornoz, J., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., Donoso, T.: Overview of EXIST 2021: sEXism Identification in Social neTworks. Procesamiento del Lenguaje Natural **67** (2021).
21. Liu, Y. et al.: RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv (2019)
22. Pires, T., Schlinger, E., Garrette, D.: How Multilingual is Multilingual BERT?. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), pp. 4996-–5001. (2019)
23. Conneau, A. et al.: Unsupervised Cross-lingual Representation Learning at Scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), pp. 8440—8451. (2020)
24. Barbieri F., Anke, L., E., Camacho-Collados, J.: XLM-T: A Multilingual Language Model Toolkit for Twitter, arXiv (2021)
25. Feng, S., Y. et al.: A Survey of Data Augmentation Approaches for NLP, arXiv (2021)
26. Paszke, A. et al.: PyTorch: An imperative style, high-performance deep learning library. In: 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), (2019)
27. Wolf, T. et al.: Transformers: State-of-the-Art Natural Language Processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP 2020), pp. 38-–45. (2020)
28. Kingma, D., P., Ba, J.: Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations (ICLR 2015), (2015)
29. Code, https://github.com/franciscorodriguez92/exist2021. Last accessed 25 May 2021
30. Transformers interpret library, https://github.com/cdpierse/transformers-interpret. Last accessed 25 May 2021