Language-based Mixture of Transformers for EXIST2024

Notebook for the EXIST Lab at CLEF 2024

Alexandru Petrescu[1], Ciprian-Octavian Truică[1,*] and Elena-Simona Apostol[1,2]

[1]National University of Science and Technology Politehnica University Bucharest, Splaiul Independenței 313, București 060042, Romania

[2]Academy of Romanian Scientists, 3 Ilfov, Bucharest, Romania

Abstract

In this paper, we propose o novel method that leverages a Mixture of Transformers (MoT) based on the language performance of each model. We employ simple, yet effective, preprocessing modules that are connected to the state-of-the-art Transformer and we compare the performance of general-purpose, task-specific, and data source-specific flavors of English and multi-language models. This novel approach manages to obtain good results for all tasks, with the best performance in soft-label evaluations rather than hard-label evaluations. We propose 3 types of mixtures that performed best on training data and we notice that they behave well against unseen data. The proposed architecture is easily up-gradable, has low resource costs, and provides good overall results in the EXIST 2024 competition.

Keywords

Mixture of Transformers, Text Classification, Learning with Disagreements, Sexism detection

1. Introduction

The following document serves as the working notes for our submission to EXIST 2024, described in [1] [2] and, representing the efforts of team Awakened. EXIST is a renowned series of scientific events and shared tasks focused on the identification of sexism in social networks. The objective of EXIST is to encompass sexism in its entirety, ranging from overt misogyny to more subtle manifestations involving implicit sexist behaviors.

For this particular event, we tackled three out of six tasks, focusing on the Natural Language Processing (NLP) challenges as the other three are similar tasks, but for Computer Vision:

- TASK 1: Sexism Identification - binary classification
- TASK 2: Source Intention - multi-class (4) classification technique, leveraging the outcome of TASK 1, identifying the nature
- TASK 3: Sexism Categorization - multi-label classification, showing the probability of each possible outcome

As most of the NLP endeavors focus on the Generative AI domain, we propose an architecture that is similar to the Mixture of Experts, but instead of using Large Language Models (LLMs) for the purpose of language generation, we use Language Models (LMs), namely Transformers, with the purpose of solving the three classification tasks. Transformers have emerged as the leading methodology for text-related operations, particularly classification problems. We take advantage of the remarkable capabilities of transformers, making use of industry-trained models facilitated by the Huggingface platform [3]. Furthermore, we fine-tune these models to optimize their performance for our particular task.

Starting with our previous work for last year's competition [4], in this article, we plan on employing a mixture of:

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

alex.petrescu@upb.ro (A. Petrescu); ciprian.truica@upb.ro (C. Truică); elena.apostol@upb.ro (E. Apostol)

https://alexpetrescu.net/ (A. Petrescu); https://sites.google.com/view/ciprian-octavian-truica (C. Truică); https://sites.google.com/view/elena-simona-apostol (E. Apostol)

0000-0002-7731-2403 (A. Petrescu); 0000-0001-7292-4462 (C. Truică); 0000-0001-6397-4951 (E. Apostol)

- general-purpose transformers
- task-specific: harmful speech detection
- data-source specific: trained on Tweets

This article is structured as follows. In Section 2, we present the current state-of-the-art methods for harmful contentment detection. In Section 3, we analyze the dataset and present the experimental setup. In Section 4, we present and discuss our results. Finally, Section 5, we draw the main conclusions of this work.

2. Related Work

The tasks proposed for this lab aim at mitigating harmful speech, more specifically sexist and offensive language, from social networks. With our work, we plan to improve the proposed approach in the previous edition of our team [4], but leverage the idea of the latest AI trend, Generative AI, namely Mixture of Experts (MoE) [5]. MoE proposes training separately a multitude of models, reducing the required resources of training a model that combines everything.

The idea of leveraging multiple simple models is not new and has been previously used for this task successfully [6], both for English and non-English tweets. Another approach that successfully uses multi-lingual transformers proposes some data augmentation techniques in the preprocessing and training pipeline [7]. An important hint that English-only embeddings might have good results in non-English tasks is provided in another working note from the previous edition [8].

Other works focus on language-independent models by training word embeddings [9], transformer embeddings [4, 10], sentence transformers [11] or document embeddings [12] for detecting online harmful content. Furthermore, in the current literature novel architectures for detecting harmful content have been proposed. These novel architectures focus on stacked deep neural networks [13] or integrating network information into their deep neural architectures [14].

Finally, the current literature also focuses on how harmful content is spread online [15, 16, 17] and how its effects can be mitigated on social platforms [18, 19, 20].

3. Experiments

3.1. Exploratory Data Analysis

To better understand the task at hand we propose a simple Exploratory Data Analysis (EDA), as we want to use a mixture of models, based on the language of the tweets. In Table 1, we observe that the proposed split of train-test $79\% - 21\%$ has the same distribution across the languages. With the balanced distribution of tweets by language $53\% - 47\%$, a mixture involving multi-lingual and English-only models makes sense and the comparisons of models will provide interesting results.

Table 1
Distribution of Tweets by Language for the Train/Test Split

DatasetSplit	Language	NumberOfItems	Percentage
Train	en	3749	47%
Train	es	4209	53%
Test	en	978	47%
Test	es	1098	53%

3.2. Experimental Setup

For our experiments, we propose a mixture of English-only and multi-lingual transformer-based models (Table 2) as we want to showcase our mixture of models architecture based on the language of the tweets (Figure 1).

The output module leverages 3 types of mixtures, in terms of output weight, for the best English and multi-language models. We consider the dominant model the English one, in case the language of the input is English, otherwise the multi-lingual one. When we present the results we will highlight the best English model like this and the best multi-lingual model like this. The leveraged mixtures are:

1. Half-Half
2. Dominant 75%
3. Dominant

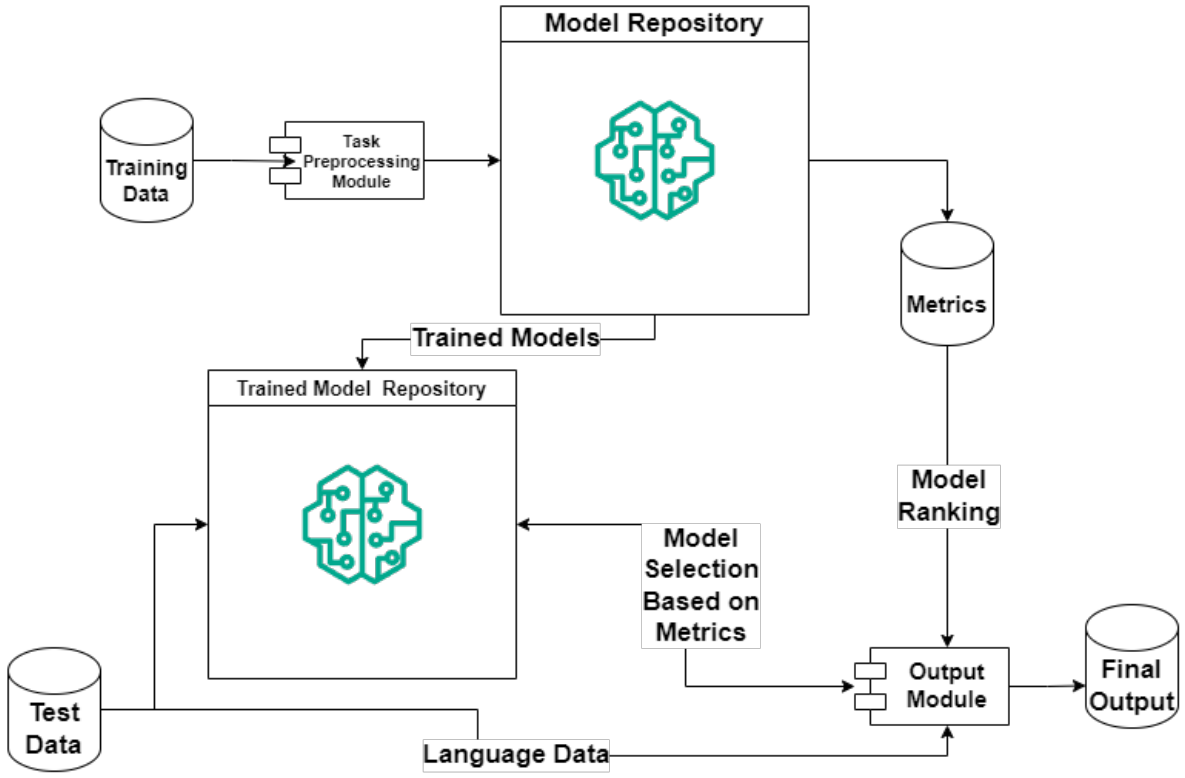

Figure 1: System Architecture

Since for this competition, Task 2 and 3 are defined to take advantage of Task 1's output, our system does the same and propagates the mixtures. This means that in Tasks 2 and Task 3, for each mixture type, the corresponding mixture from Task 1 is used.

Table 2
The proposed models for our experiments

ModelName	IsMultiLingual
twitter-roberta [21] [22]	No
twitter-xlm-roberta-base-sentiment-multilingual [21]	Yes
twitter-xlm-roberta-base-sentiment [23]	Yes
bert-toxic-comment-classification [24]	No
distilbert-uncased-english [25]	No
distilbert-base-multilingual-cased-sentiments [26]	Yes
MiniLM-L12-H384 [27]	No
xlm-roberta [28]	Yes
roberta-hate-speech-dynabench-r4 [29]	No

For all the tasks, we use early stopping with 3 epochs of tolerance and the following hyper-parameters, obtained while training the best model strategy:

- $learning\_rate = 2e^{-5}$
- $per\_device\_train\_batch\_size = 32$
- $per\_device\_eval\_batch\_size = 32$
- $weight\_decay = 0.01$
- $max\_epochs = 50$

The metrics used in the competition, for which the engine will be optimized, are ICM-Hard, ICM-Hard Norm, F1, Cross Entropy, Majority class, Minority class, and Oracle most voted. To provide models that perform well we are using F1 for Tasks 1 and 2 and for Task 3 we are using a custom Mean Squared Error.

As for the hyper-parameter tuning, each model is optimized as it would be handling the task alone, for the current implementation.

3.2.1. Task 1

The first task is a binary classification one. The system has to decide whether or not a given tweet contains sexist expressions or behaviors. The dataset is annotated by multiple evaluators, each providing their own label. To obtain only one label for each tweet, i.e., *'YES'* or *'NO'*, we take the majority label.

In Table 3, we observe that the best-performing models for this task are the ones fine-tuned on Twitter data. As mentioned in the previous section, we are going to use a mixture of the best English-based model and the best multi-lingual model.

Table 3
General Metrics for the Proposed Models on Task 1

ModelName	Epoch	F1	Loss	Train(s)	Eval(s)
twitter-roberta	4	0.7789	0.4863	1463	5
twitter-xlm-roberta-base-sentiment-multilingual	3	0.7665	0.4670	7039	159
twitter-xlm-roberta-base-sentiment	3	0.7482	0.4902	6373	146
bert-toxic-comment-classification	4	0.7463	0.5211	6969	117
distilbert-uncased-english	4	0.7406	0.5348	2918	3
distilbert-base-multilingual-cased-sentiments	4	0.7379	0.5123	4407	3
MiniLM-L12-H384	5	0.7338	0.5059	296	3
xlm-roberta	4	0.7327	0.5520	1834	9
roberta-hate-speech-dynabench-r4	3	0.7126	0.5220	6080	154

We notice that the models are close when it comes to performance, all are in the range of $71\% - 79\%$, but when it comes to the resources used there is a meaningful difference. The least resources are used by MiniLM [27], which is a super-pruned version of the regular LMs. The most resource-intensive models are the ones that are trained in multiple extra iterations over the regular LMs, namely the multi-lingual tranformers. The base for the multi-lingual transformers is XLM Roberta. Each is further trained on platform-specific data, i.e., tweets from Twitter, and task-specific data, i.e., harmful speech.

3.2.2. Task 2

The second task is multi-class classification, namely "Source Intention". Building on top of the first one, it aims to categorize the message according to the intention of the author. This provides insights into the role played by social networks in the emission and dissemination of sexist messages. To unify the results, we use the same approach as for Task 1, a majority vote with equal weight for the labels. Furthermore, we are augmenting the output by leveraging the output from Task 1, working as a *'YES'* or *'NO'* filter that tells us if the model needs to be run on the input or not.

For this problem, we observe that the range of results is wider (Table 4), i.e., $46\% - 61\%$. The models perform significantly worse than they previously did, but this is expected as the output of Task 1 is also leveraged.

As expected, the specialized models are outperforming the others. Moreover, the English model, which is solely focused on the task at hand rather than the data, has the overall best performance by a small margin. Resource-wise, the behavior is not reflected on the macro level as it was for Task 1. As such, MiniLM, despite training for the most epochs among the smaller models, is not the least demanding. However, it remains the most efficient model per iteration.

Table 4
General Metrics for the Proposed Models on Task 2

ModelName	Epoch	F1	Loss	Train(s)	Eval(s)
twitter-xlm-roberta-base-sentiment	5	0.6090	0.8623	760	4
xlm-roberta	7	0.6064	0.9265	1107	4
roberta-hate-speech-dynabench-r4	5	0.6063	0.8843	515	2
twitter-roberta	5	0.5822	0.8790	518	2
twitter-xlm-roberta-base-sentiment-multilingual	4	0.5525	0.8631	625	4
MiniLM-L12-H384	10	0.5395	0.9431	482	1
distilbert-uncased-english	5	0.5333	0.9226	115	1
bert-toxic-comment-classification	4	0.5021	0.9227	176	2
distilbert-base-multilingual-cased-sentiments	4	0.4657	0.9439	126	1

3.2.3. Task 3

Task 3 is a multi-label classification focusing on identifying different sexism categories for each tweet that was labeled as sexist by Task 1. Unlike Task 1, tweets have multiple sexist labels. Thus, our proposed approach computes a probability for each label considering the number of annotations, using an equal weight for each annotation. As the metrics are custom, the loss is $1/Metric$, and we did not need to represent it in Table 5. The custom Mean Square Error (CustomMSE) is adapted to the way we build the probabilities of each class.

We observe an almost perfect mirror of what happened before (Table 5), with the best performing models being the data and task-specific ones. For the resources side, we notice that this time MiniLM trained more than twice the number of epochs that the others.

Table 5
General Metrics for the Proposed Models on Task 3

ModelName	Epoch	CustomMSE	Train(s)	Eval(s)
twitter-xlm-roberta-base-sentiment-multilingual	7	32.6568	1154	5
twitter-xlm-roberta-base-sentiment	7	32.0696	1089	5
xlm-roberta	7	31.0823	1060	2
twitter-roberta	7	30.2483	664	2
distilbert-uncased-english	7	29.6718	170	2
roberta-hate-speech-dynabench-r4	8	29.5671	826	3
distilbert-base-multilingual-cased-sentiments	7	29.2608	233	2
bert-toxic-comment-classification	8	29.2145	396	3
MiniLM-L12-H384	18	27.8630	862	1

4. Results

Table 6 presents the official results from the leaderboard. For a more comprehensive analysis, please refer to the Results chapter available on the official site. We are showcasing only the best ranking, as most of the submissions are one after another in the rankings with a maximum drift of 3 places.

We notice that the best overall mixture is 2, $DOMINANT - 75\%$, and the least is mixture 1, $HALF - HALF$. The best-performing outputs are on the English tasks for the soft evaluation rather

Table 6
Ranking in EXIST2024 competition

Task	EvalType	BestMixture	BestRank	TotalSystems
1	Soft-Soft-ALL	2	10	40
1	Hard-Hard-ALL	3	20	70
1	Soft-Soft-ES	3	16	40
1	Hard-Hard-ES	3	21	66
1	Soft-Soft-EN	1	5	40
1	Hard-Hard-EN	3	12	68
2	Soft-Soft-ALL	2	9	35
2	Hard-Hard-ALL	2	12	46
2	Soft-Soft-ES	2	11	35
2	Hard-Hard-ES	2	14	46
2	Soft-Soft-EN	2	11	35
2	Hard-Hard-EN	2	7	46
3	Soft-Soft-ALL	2	9	33
3	Hard-Hard-ALL	2	6	34
3	Soft-Soft-ES	2	10	33
3	Hard-Hard-ES	2	9	34
3	Soft-Soft-EN	3	8	33
3	Hard-Hard-EN	2	6	34

than the hard one. One interesting aspect is that we obtained the lowest performance for Task 1, but the other two that are leveraging its output behave better, with a slight margin. Another interesting aspect is that Task 3, the one that leverages the custom metric, has the best results out of all tasks, with consistent placement.

One thing to notice is the difference between the soft and the hard evaluation, for all language splits, where for Task 1 the difference is quite significant and for the other 2 not that much, considering that in most cases the outputs of each team were one after another and each team had 3 possible outputs that can be submitted.

To conclude the Mixture of Transformers provides promising results with good resource requirements, with the second proposed mixture, $DOMINANT - 75\%$, performing on average the best, with the difference in performance between them being not that significant.

5. Conclusions and future directions

We notice a similar performance as we did in the previous iteration, namely in [4], which is slightly fixed by the Mixture of Transformers:

1. The models yield better results for the soft evaluation, meaning we can adjust the tolerance to better improve the hard evaluation.
2. The models behave better on the English data, which means that we have to either find better models specialized in other languages or fine-tune the multi-language ones with more data.

One thing that we did not tackle, but we previously mentioned, is experimenting with the way we weigh each label, based on the meta-data of the annotator, but the literature has mixed views on this.

Another interesting approach is to consider a dynamic number of Transformers, for each language, based on performance, as we observe that sometimes the performance is close for multiple models.

Acknowledgment

This work is supported in part by

- The German Academic Exchange Service (DAAD) through the project "iTracing: Automatic Misinformation Fact-Checking" (DAAD grant no. 91809005).
- The Academy of Romanian Scientists through the funding of project "SCAN-NEWS: Smart system for deteCting And mitigatiNg misinformation and fake news in social media" (AOȘR-TEAMS-III).

References

[1] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[2] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes (Extended Overview), in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[3] 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, Huggingface 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, 2020, pp. 38–45. doi:10.18653/v1/2020.emnlp-demos.6.

[4] A. Petrescu, Leveraging MiniLMv2 Pipelines for EXIST2023, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 1037–1043.

[5] O. Sanseviero, L. Tunstall, P. Schmid, S. Mangrulkar, Y. Belkada, P. Cuenca, Mixture of experts explained, 2023. URL: https://huggingface.co/blog/moe.

[6] C. Jhakal, K. Singal, M. Suri, D. Chaudhary, B. Kumar, I. Gorton, Detection of sexism on social media with multiple simple transformers, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), volume 3497 of *CEUR Workshop Proceedings*, 2023, pp. 959–966.

[7] H. Mohammadi, A. Giachanou, A. Bagheri, Towards robust online sexism detection: A multimodel approach with bert, xlm-roberta, and distilbert for EXIST 2023 tasks, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), volume 3497 of *CEUR Workshop Proceedings*, 2023, pp. 1000–1011.

[8] A. Sanchez-Urbina, H. Gómez-Adorno, G. Bel-Enguix, V. Rodríguez-Figueroa, A. Monge-Barrera, Iimasgil_nlp@exist2023: Unveiling sexism on twitter with fine-tuned transformers, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 1067–1082.

[9] V.-I. Ilie, C.-O. Truică, E.-S. Apostol, A. Paschke, Context-Aware Misinformation Detection: A Benchmark of Deep Learning Architectures Using Word Embeddings, IEEE Access 9 (2021) 162122–162146. doi:10.1109/access.2021.3132502.

[10] C.-O. Truică, E.-S. Apostol, MisRoBÆRTa: Transformers versus Misinformation, Mathematics 10 (2022) 1–25(569). doi:10.3390/math10040569.

[11] C.-O. Truică, E.-S. Apostol, A. Paschke, Awakened at CheckThat! 2022: fake news detection using BiLSTM and sentence transformer, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF2022), 2022, pp. 749–757.

[12] C.-O. Truică, E.-S. Apostol, It's All in the Embedding! Fake News Detection Using Document Embeddings, Mathematics 11 (2023) 508. doi:10.3390/math11030508.

[13] E.-S. Apostol, C.-O. Truică, A. Paschke, Contcommrtd: A distributed content-based misinformation-

aware community detection system for real-time disaster reporting, IEEE Transactions on Knowledge and Data Engineering (2024) 1–12. doi:10.1109/tkde.2024.3417232.

[14] C.-O. Truică, E.-S. Apostol, P. Karras, DANES: Deep Neural Network Ensemble Architecture for Social and Textual Context-aware Fake News Detection, Knowledge-Based Systems 294 (2024) 1–13(111715). doi:10.1016/j.knosys.2024.111715.

[15] A. Petrescu, C.-O. Truică, E.-S. Apostol, Sentiment Analysis of Events in Social Media, in: 2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP), IEEE, 2019, pp. 143–149. doi:10.1109/iccp48234.2019.8959677.

[16] A. Petrescu, C.-O. Truică, E.-S. Apostol, A. Paschke, EDSA-Ensemble: an Event Detection Sentiment Analysis Ensemble Architecture, 2023. arXiv:2301.12805.

[17] C.-O. Truică, E.-S. Apostol, T. Ștefu, P. Karras, A Deep Learning Architecture for Audience Interest Prediction of News Topic on Social Media, in: International Conference on Extending Database Technology (EDBT2021), 2021, pp. 588–599. doi:10.5441/002/EDBT.2021.69.

[18] A. Petrescu, C.-O. Truică, E.-S. Apostol, P. Karras, Sparse Shield: Social Network Immunization vs. Harmful Speech, in: ACM International Conference on Information and Knowledge Management (CIKM2021), ACM, 2021, pp. 1426–1436. doi:10.1145/3459637.3482481.

[19] C.-O. Truică, E.-S. Apostol, R.-C. Nicolescu, P. Karras, MCWDST: a Minimum-Cost Weighted Directed Spanning Tree Algorithm for Real-Time Fake News Mitigation in Social Media, IEEE Access 11 (2023) 125861–25873. doi:10.1109/ACCESS.2023.3331220.

[20] E.-S. Apostol, Özgur Coban, C.-O. Truică, Contain: A community-based algorithm for network immunization, Engineering Science and Technology, an International Journal 55 (2024) 1–10(101728). doi:10.1016/j.jestch.2024.101728.

[21] J. Camacho-Collados, K. Rezaee, T. Riahi, A. Ushio, D. Loureiro, D. Antypas, J. Boisson, L. Espinosa Anke, F. Liu, E. Martínez Cámara, Tweetnlp: Cutting-edge natural language processing for social media, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, 2022, pp. 38–49. doi:10.18653/v1/2022.emnlp-demos.5.

[22] D. Loureiro, F. Barbieri, L. Neves, L. Espinosa Anke, J. Camacho-collados, TimeLMs: Diachronic language models from Twitter, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 251–260. doi:10.18653/v1/2022.acl-demo.25.

[23] F. Barbieri, L. Espinosa Anke, J. Camacho-Collados, XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 258–266.

[24] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 2019. doi:10.18653/v1/n19-1423.

[25] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. arXiv:1910.01108.

[26] M. Laurer, W. van Atteveldt, A. Casas, K. Welbers, Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli, Political Analysis 32 (2024) 84–100. doi:10.1017/pan.2023.20.

[27] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020. arXiv:2002.10957.

[28] 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, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL, 2020, pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747.

[29] B. Vidgen, T. Thrush, Z. Waseem, D. Kiela, Learning from the worst: Dynamically generated datasets to improve online hate detection, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), ACL, 2021, pp. 1667–1682. doi:10.18653/v1/2021.acl-long.132.