Awakened at EXIST2025: Adaptive Mixture of Transformers

Notebook for the EXIST2025 Lab at CLEF 2025

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

This paper presents an adaptive Mixture of Transformers architecture developed for the EXIST 2025 Lab, targeting the detection of sexist content in social media text. The proposed system combines nine transformer-based models—spanning both English-specific and multilingual variants—each specialized by language, platform, or task. A dynamic weighting mechanism automatically adjusts the contribution of each model in the ensemble based on the detected language and performance metrics, enabling robust and context-aware classification across diverse linguistic settings. Experimental results demonstrate that the 2025 architecture achieves competitive performance compared to previous years, surpassing the 2023 and 2024 iterations in several Subtasks and delivering notable improvements in cross-lingual and task-specific detection. However, results for certain Subtasks indicate areas for further optimization. The findings highlight the effectiveness of adaptive model selection and weighting in ensemble architectures for harmful content detection and suggest promising directions for future research in multilingual and context-sensitive text classification.

Keywords

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

1. Introduction

The increase of harmful content on social media platforms presents a significant challenge for both researchers and practitioners, particularly when it comes to the detection and mitigation of sexist material. Addressing these issues requires robust, adaptable, and multilingual solutions capable of operating effectively across diverse linguistic and contextual settings. Building on previous work in the field, this paper introduces an enhanced approach for the EXIST 2025 Lab [1, 2], focusing on the detection of sexist content in textual data using an adaptive mixture of transformer-based models.

Our objective is to improve upon earlier iterations of the Mixture of Transformers [3] architecture by overcoming identified limitations and leveraging recent advancements in transformer models. While the EXIST Lab encompasses multiple content types—including text, images, and videos—this work concentrates exclusively on text classification, reflecting both the prevalence of textual interactions on social networks and the unique challenges posed by language-based harmful content.

The proposed system integrates nine distinct transformer models, including both English-specific and multilingual variants, each tailored to specific languages, tasks, or social media platforms. Central to our approach is a dynamic weighting mechanism that automatically adjusts the contribution of each model within the ensemble based on the detected language and relevant performance metrics. This adaptive strategy ensures optimal performance across a wide range of linguistic scenarios and enhances the system's ability to detect and classify sexist content accurately.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

[*] Corresponding author.

alex.petrescu@upb.ro (A. Petrescu); elena.apostol@upb.ro (E. Apostol); ciprian.truica@upb.ro (C. Truică)
https://alexpetrescu.net/ (A. Petrescu); https://sites.google.com/view/elena-simona-apostol (E. Apostol); https://sites.google.com/view/ciprian-octavian-truica (C. Truică)
0000-0002-7731-2403 (A. Petrescu); 0000-0001-6397-4951 (E. Apostol); 0000-0001-7292-4462 (C. Truică)

By advancing the state of the art in mixture-of-experts architectures and addressing the complex problem of learning with disagreements, this paper aims to contribute effective tools for reducing harmful content and fostering safer online communities.

2. Related Work

In recent years, researchers have addressed harmful content detection using three primary approaches. The most prevalent method involves combining word and transformer-based embeddings with deep learning techniques to classify textual data [4, 5, 6, 7, 3]. Another research direction focuses on enriching contextual understanding by incorporating metadata, such as social context [8] or tracking how information spreads within a network [9]. Additionally, emerging strategies advocate for the application of network immunization techniques to stop the spread and dissemination of harmful content [10, 11, 12, 13, 14, 15]. Furthermore, holistic systems and architectures have also been developed to monitor and analyze social media content in real-time to detect and stop the spread of harmful content [16, 17].

The objectives of this lab focus on reducing harmful content on social networks and not only with a particular focus on addressing sexist material. Our goal is to enhance the approach developed by our team in previous editions, [18] and [3], by addressing some of the shortcomings identified in the earlier Mixture of Transformers architecture. While the lab proposes 3 types of content (text, images and videos), we will be focusing on text.

As the current literature also focuses on how harmful content and its spread online [10, 12, 11], leveraging Mixture of Experts architectures to pretrained BERT models [19], we can later start a discussion on how its effects can be mitigated on social platforms [13, 14, 15].

Our system architecture, showcased in Figure 1, is built to effectively detect sexist content in social media textual content by utilizing a mixture of transformer-based models combined with another learner capable of automatically adjusting the contribution of each transformer within the ensemble. The architecture integrates a total of nine distinct transformer models, including four that support multiple languages, each specialized in either the language, task, or platform from which the content comes. The central concept is to dynamically select and weight the models based on the detected language of the input tweet and relevant performance metrics, ensuring optimal results across various linguistic settings.

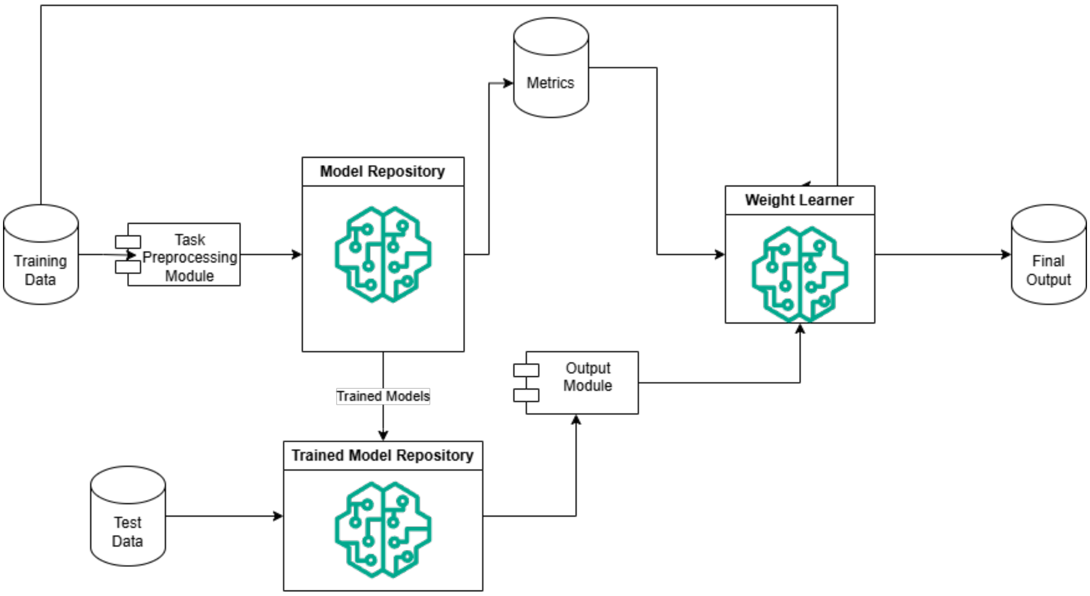

Figure 1: The Proposed Architecture

In order to have proper experiments, we have chosen the same transformer model repository as the one in the 2024 architecture [3], provided by Hugging Face. Our selection of English-specific models for these tasks includes:

- **twitter-roberta**: a model pre-trained on Twitter data, ideal for social media language.
- **bert-toxic-comment-classification**: a BERT-based model fine-tuned for detecting toxic content.
- **distilbert-uncased-english**: a smaller, faster, and more efficient version of BERT.
- **MiniLM-L12-H384**: a highly compact but powerful language model.
- **roberta-hate-speech-dynabench-r4**: a RoBERTa variant specifically trained for hate speech detection.

For our system's multilingual functionality, we selected the following transformer models:

- **twitter-xlm-roberta-base-sentiment**: an XLM-RoBERTa variant fine-tuned for sentiment analysis on Twitter data across multiple languages.
- **twitter-xlm-roberta-base-sentiment-multilingual**: this model is similar to the above but fine-tuned on a broader, more diverse multilingual sentiment dataset.
- **distilbert-base-multilingual-cased-sentiments**: a more compact and efficient version of BERT, pre-trained on a vast multilingual corpus and fine-tuned for sentiments.
- **xlm-roberta**: a multilingual model pre-trained on an enormous dataset covering over 100 languages.

3. Experiments

Building upon the mixture architectures introduced in our 2024 work [3], the 2025 Lab implementation employs three transformer configurations with automated weight optimization:

- **Full Ensemble** - Configuration 1: Uses all nine transformer models, with weights dynamically adjusted by the learning algorithm.
- **Top Performers** - Configuration 2: Combines specialized leaders from distinct categories: twitter-roberta (platform-optimized), twitter-xlm-roberta-base-sentiment-multilingual (multilingual analysis), and bert-toxic-comment-classification (task-specific detection).
- **Best and Worst - Roberta** - Configuration 3: Averaged contributions for Subtasks 1.1 and 1.2 for Roberta variations and picked the best 1 and worst 3 to add some noise, resulting in: twitter-roberta, twitter-xlm-roberta-base-sentiment, roberta-hate-speech-dynabench-r4, and twitter-xlm-roberta-base-sentiment-multilingual.

Table 1
Subtask 1.1 and 1.2 Ensemble Model Contributions

Model Name	Imp - 1	% - 1	Imp - 2	% - 2	AVG Imp	AVG %
twitter-roberta	**1.0000**	**0.3813**	**0.3783**	**0.1272**	**0.6892**	**0.2543**
xlm-roberta	0.0096	0.0037	1.0000	0.3363	0.5048	0.1700
distilbert-base-multilingual-cased-sentiments	0.7615	0.2904	0.0352	0.0118	0.3983	0.1511
MiniLM-L12-H384	0.0612	0.0233	0.6425	0.2161	0.3519	0.1197
bert-toxic-comment-classification	0.5732	0.2186	0.0326	0.0110	0.3029	0.1148
twitter-xlm-roberta-base-sentiment	**0.0107**	**0.0041**	**0.4797**	**0.1613**	**0.2452**	**0.0827**
distilbert-uncased-english	0.1742	0.0664	0.1095	0.0368	0.1418	0.0516
roberta-hate-speech-dynabench-r4	**0.0096**	**0.0037**	**0.2385**	**0.0802**	**0.1241**	**0.0419**
twitter-xlm-roberta-base-sentiment-multilingual	**0.0226**	**0.0086**	**0.0569**	**0.0191**	**0.0398**	**0.0139**

In Table 1 we showcase the contribution of each model in the ensemble for the main configuration (Full Ensemble) for Subtasks 1.1 and 1.2. We analyze the importance of each model in the ensemble (Imp - 1 for Subtask 1.1 and Imp - 2 for Subtask 1.2) and the Contribution to the output Percentage (% -

1 for Subtask 1.1 and % - 2 for Subtask 1.2). Based on these results, we have derived configurations 2 and 3. In the considerations for the ensemble weights, we have taken only the single-label classification Subtasks 1.1 and 1.2 and not Subtask 1.3, which is multi-label classification, for this version of the system. Each sub-task and configuration employs its own learning module, with the same learning strategy: maximize the hard-label metric. This dynamic weighting strategy enables context-aware prioritization of models, improving both detection accuracy and cross-lingual robustness.

For our experiments, we propose a mixture of English-only and multi-lingual transformer-based models, presented in Table 2, as we want to showcase our mixture of models architecture, as seen in Figure 1, based on the language of the tweets. The output module leverages 3 types of mixtures, described above. Since the structure of this competition is that Subtasks 1.2 and 1.3 leverage the output of Subtask 1.1, our system does the same, so we propagate the mixtures, meaning that for Subtasks 1.2 and 1.3, for each mixture type, the corresponding mixture from Subtask 1.1 is used. For all the learning tasks, we have used an early stop with 3 epochs of tolerance, the best model strategy, and the following hyperparameters:

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

Table 2
Models used in our experiments

Hugging Face Model	**IsMultiLingual**
cardiffnlp/twitter-roberta-base-sentiment-latest	No
cardiffnlp/twitter-xlm-roberta-base-sentiment-multilingual	**Yes**
cardiffnlp/twitter-xlm-roberta-base-sentiment	**Yes**
JungleLee/bert-toxic-comment-classification	No
distilbert/distilbert-base-uncased-finetuned-sst-2-english	No
lxyuan/distilbert-base-multilingual-cased-sentiments-student	**Yes**
microsoft/Multilingual-MiniLM-L12-H384	No
papluca/xlm-roberta-base-language-detection	**Yes**
facebook/roberta-hate-speech-dynabench-r4-target	No

According to the official competition guidelines, the evaluation metrics include ICM-Hard, ICM-Hard Norm, F1-Score, Cross Entropy, Majority class, Minority class, and Oracle most voted. Our system is specifically optimized for these metrics. For Subtasks 1.1 and 1.2, we focus on maximizing the F1-Score [20], while for Subtask 1.3, we employ a custom Mean Squared Error as the primary metric.

Regarding hyperparameter tuning, each model is individually optimized as if it were solving the Subtask independently in the current setup and then the Weight Adjuster learns the weights from the outputs of each individual transformer based on the configuration.

4. Discussion

The proposed solution achieved strong leaderboard results, and we are analyzing its performance in comparison to the 2024 [3] and 2023 [18] approaches. Table 3 presents the overall evaluation metrics—both soft and hard—for all languages.

For Subtask 1.1, the proposed 2025 architecture obtains better results than the 2023 architecture, but falls short to the 2024 one, both in soft and hard evaluations, with a bigger difference in the soft one. Surprisingly, the best behaving configuration is number 3 (Best and Worst - Roberta), followed by number 1 shortly (Full Ensemble), and with a bit of a difference by number 2 (Top Performers).

Table 3
EXIST 2023 vs 2024 vs 2025 architectures, Soft Label Evaluation, All Languages

Run	Subtask	ICM-Soft	ICM-Soft Norm	Cross Entropy	ICM-Hard	ICM-Hard Norm	F1-Score
2023	1	0.3214	0.5482	1.1709	0.4021	0.6222	0.73
2024_2	1.1	**0.7196**	**0.6154**	0.8106	0.5124	0.7575	0.762
2024_3	1.1	0.6909	0.6108	0.8542	**0.5196**	**0.7611**	**0.765**
2024_1	1.1	0.6663	0.6068	**0.8037**	0.4984	0.7505	0.758
2025_3	1.1	0.5328	0.5854	0.9084	0.4562	0.7293	0.75
2025_1	1.1	0.531	0.5851	0.9531	0.455	0.7287	0.749
2025_2	1.1	0.4074	0.5653	0.9645	0.435	0.7186	0.739
2023	1.2	-3.1765	**0.7604**	3.205	-0.1481	**0.6407**	0.428
2024_2	1.2	**-2.0091**	0.3381	**3.0835**	**0.1812**	0.5589	**0.483**
2024_1	1.2	-2.0365	0.3359	3.1429	0.1487	0.5483	0.475
2024_3	1.2	-2.1502	0.3268	3.0908	0.1306	0.5425	0.469
2025_2	1.2	-7.8903	0	6.9973	-2.2184	0	0.22
2025_1	1.2	-9.0253	0	5.2347	-2.2016	0	0.233
2025_3	1.2	-18.446	0	17.426	-2.1971	0	0.236
2023	1.3	-4.2139	**0.7538**	N/A	-0.1948	**0.5555**	0.482
2024_2	1.3	-4.0748	0.2848	N/A	-0.0042	0.499	0.483
2024_3	1.3	-4.0786	0.2846	N/A	-0.0115	0.4973	0.48
2024_1	1.3	-4.1845	0.279	N/A	-0.0427	0.4901	0.474
2025_2	1.3	**-3.7444**	0.3023	N/A	0.1078	0.525	0.553
2025_1	1.3	-3.7539	0.3018	N/A	**0.1491**	0.5346	**0.563**
2025_3	1.3	-3.9428	0.2918	N/A	0.1195	0.5277	0.552

For Subtask 1.2, the results are worse than both 2024 and 2023 systems in all metrics by a considerable amount, for both soft and hard evaluations, which means that we have something to improve on and consider a main point of interest. Here, the best behaving configuration is number 2 (Top Performers), followed by a small difference by number 1 (Full Ensemble), and then, by a big difference by number 3 (Best and Worst - Roberta).

For Subtask 1.3, we have managed to improve our results over the 2023 and 204 iterations, by some margin, both in the soft and hard evaluations. The best behaving configuration is also number 2 (Top Performers), followed by a small difference by number 1 (Full Ensemble), and then again by a small difference by number 3 (Best and Worst - Roberta).

5. Conclusions and future directions

The adaptive Mixture of Transformers architecture developed for EXIST 2025 demonstrates the value of leveraging diverse transformer models over certain categories, such as multilingual, task-specific, and platform-specific. By learning the weights for the transformer ensemble according to the provided configurations, the system achieves robust and context-aware performance across multiple languages and evaluation settings. Experimental results show that the 2025 solution performs competitively on the leaderboard, surpassing the 2023 approach in several Subtasks and offering improvements in Subtask 1.3 over both previous iterations. However, Subtask 1.2 results indicate areas where further optimization is needed, particularly in both soft and hard evaluation metrics.

The analysis of different mixture configurations reveals that adaptive selection and weighting of transformer models can significantly influence performance, with certain configurations excelling in specific Subtasks. The architecture remains resource-efficient and easily upgradable, supporting ongoing experimentation and refinement.

For future work, by leveraging Subtask 1.1 findings, we have noticed that adding noise to the ensemble provided an interesting result, which needs to be analyzed over other datasets in order to confirm this behavior.

From Subtask 1.2 findings, we can also apply multiple strategies such as better metrics for multi-class classification, not focusing on only the hard-label evaluation, and choosing better models for the model repository.

As for the findings from Subtask 1.3, there can also be an ensemble custom-tailored with its results, on the soft-label evaluation, compared to the hard-label evaluations that have been considered in the current iteration of the system.

Acknowledgments

The research presented in this paper was supported in part by (1) The Academy of Romanian Scientists through the funding of the project “SCAN-NEWS: Smart system for deteCting And mitigatiNg misinformation and fake news in social media” (AOȘR-TEAMS-III); (2) The Academy of Romanian Scientists through the funding of the project “NetGuardAI: Intelligent system for harmful content detection and immunization on social networks” (AOȘR-TEAMS-IV); and (3) the National University of Science and Technology, POLITEHNICA Bucharest, through the PubArt program.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of exist 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos. experimental ir meets multilinguality, multimodality, and interaction., in: Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[2] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of exist 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos (extended overview)., in: CLEF 2025 Working Notes, 2025.

[3] A. Petrescu, C.-O. Truică, E.-S. Apostol, Language-based Mixture of Transformers for EXIST2024, in: Working Notes of the Conference and Labs of the Evaluation Forum, volume 3740 of *CEUR Workshop Proceedings*, 2024, pp. 1157–1164.

[4] 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.

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

[6] 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, 2022, pp. 749–757.

[7] C.-O. Truică, E.-S. Apostol, It’s all in the Embedding! Fake News Detection using Document Embeddings, Mathematics 11 (2023) 1–29(508). doi:10.3390/math11030508.

[8] 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:https://doi.org/10.1016/j.knosys.2024.111715.

[9] C.-O. Truică, E.-S. Apostol, M. Marogel, A. Paschke, GETAE: Graph Information Enhanced Deep Neural NeTwork Ensemble ArchitecturE for fake news detection, Expert Systems with Applications 275 (2025) 126984. doi:10.1016/j.eswa.2025.126984.

[10] 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.

[11] A. Petrescu, C.-O. Truică, E.-S. Apostol, A. Paschke, EDSA-Ensemble: an Event Detection Sentiment Analysis Ensemble Architecture, IEEE Transactions on Affective Computing (2024) 1–18. doi:10.1109/TAFFC.2024.3434355.

[12] 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.

[13] 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.

[14] 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) 125861–125873. doi:10.1109/ACCESS.2023.3331220.

[15] 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:https://doi.org/10.1016/j.jestch.2024.101728.

[16] 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.

[17] C.-O. Truică, A.-T. Constantinescu, E.-S. Apostol, StopHC: A Harmful Content Detection and Mitigation Architecture for Social Media Platforms, in: IEEE International Conference on Intelligent Computer Communication and Processing (ICCP 2024), 2024, pp. 1–5. doi:10.1109/ICCP63557.2024.10793051.

[18] 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.

[19] L. Hallee, R. Kapur, A. Patel, J. P. Gleghorn, B. Khomtchouk, Contrastive learning and mixture of experts enables precise vector embeddings, 2024. URL: https://arxiv.org/abs/2401.15713. arXiv:2401.15713.

[20] C.-O. Truică, C. A. Leordeanu, Classification of an imbalanced data set using decision tree algorithms, Univiversity Politechnica of Bucharest Scientific Bulletin - Series C Electrical Engineering and Computer Science 79 (2017) 69–84.