Multilingual Sexism Identification via Fusion of Large Language Models

Sahrish Khan[1,*], Gabriele Pergola[1] and Arshad Jhumka[3]

[1]Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK
[3]School of Computing, University of Leeds, Leeds LS2 9JT, UK

Abstract

The pervasive presence of sexist content on social media platforms not only perpetuates harmful stereotypes but also fosters environments that can be exclusionary and hostile, especially towards women. Such content, which often targets people of a specific gender, i.e., sexist content, requests platforms to enhance their monitoring and policing efforts. Yet, policing such content is challenging for many reasons, including the volume of messages to check and the context of the content. Consequently, several studies have been conducted to automatically detect sexist language on social media, focusing on its identification and classification. However, variations in detection accuracy can depend on the differences in architecture, training strategies, and data of existing models, [1, 2, 3], leading to potential variances in detection accuracy. This variability, further influenced by the types of messages and input prompts, motivates our exploration into the fusion of multiple Large Language Models (LLMs). As part of EXIST Task 1, which focuses on sexism identification in multilingual contexts, we introduce two novel approaches: the *Dual-Transformer Fusion Network (DTFN)* and the *Multimodel Fusion Ensemble (MFE)*. These methods utilize fusion and ensemble learning techniques to enhance detection accuracy across multilingual datasets. Our extensive experimental evaluation during the EXIST 2024 competition demonstrates that these methodologies significantly outperform existing models, with MFE and DTFN ranking 1st and 2nd, respectively, in the English segment, and 4th and 13th in the combined English and Spanish segments of the official leaderboard.

Keywords

Social Media, Sexism Detection, Ensemble, Transformer, Multilingual, Large Language models, EXIST Task 1

1. Introduction

The proliferation of social media platforms has fundamentally transformed how individuals communicate. However, these platforms have also become arenas for problematic interactions, including the dissemination of sexist content. This content not only perpetuates harmful stereotypes but also fosters an online environment that can be hostile and exclusionary, particularly towards women. The urgency to address this issue is underscored by the growing body of research indicating the growing exposure to sexist language and its profound impacts. Despite the clear need to mitigate this problem, the task of detecting sexist content online presents substantial challenges. Sexist language is not uniformly explicit; it often involves subtle cues and context-dependent expressions.

Addressing these challenges requires leveraging flexible computational methods and approaches that can understand and interpret the complexities of language used in these settings. Large Language Models (LLMs), which are pre-trained on vast corpora and fine-tuned for specific tasks, are promising solutions. However, while individual LLMs offer robust linguistic insights, they also have inherent limitations when applied to specific tasks or domains such as detecting sexist or harmful content. Each existing model may interpret nuances differently based on its architecture, training strategy and data [1, 2, 3, 4], leading to potential variances in detection accuracy.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding author.
†These authors contributed equally.
sahrish.khan@warwick.ac.uk (S. Khan); gabriele.pergola.1@warwick.ac.uk (G. Pergola); arshad.jhumka@leeds.ac.uk (A. Jhumka)
https://www.dcs.warwick.ac.uk/~u2149613/ (S. Khan); https://www.dcs.warwick.ac.uk/~u1898418/ (G. Pergola); https://eps.leeds.ac.uk/computing/staff/9540/professor-arshad-jhumka (A. Jhumka)
0000-0002-7347-2522 (G. Pergola); 0000-0003-0540-2845 (A. Jhumka)

This variability depends on the types of messages as well as on the input prompts, and it motivates our exploration for fusing multiple LLMs. Our research, conducted as part of the EXIST 2024 (sEXism Identification in Social Networks)[5, 6] shared task 1, which focuses on enhancing automated sexism detection. In this paper, we present two novel methodologies leveraging neural language models for sexism identification: the *Dual-Transformer Fusion Network (DTFN)* and the *Multimodel Fusion Ensemble (MFE)*. These approaches utilize fusion and ensemble learning techniques to enhance detection accuracy across multilingual datasets, specifically evaluated using the EXIST 2024 dataset for both English and Spanish contexts.

In particular, the DTFN is a simple yet effective approach that integrates the outputs from two distinct transformers, i.e., RoBERTa[1] and DeBERTa [2], and fuses them via a fully connected layer. We posit that by concatenating their outputs, the DTFN captures a more comprehensive understanding of the textual data. We further expand this concept by introducing the MFE approach, which applies a majority voting mechanism among multiple models to exploit their collective capabilities for better generalization across diverse linguistic contexts. Ensemble methods, such as MFE, have shown to enhance model performance by mitigating individual model weaknesses and reducing the variance of predictions [7]. By incorporating a diverse set of models like RoBERTa-Large, DeBERTa-V3-Large, Mistral-7b [1, 2, 3], and DTFN, the MFE approach provides a more robust and accurate detection methodology.

The effectiveness of these methods was evaluated in the EXIST 2024 competition - Task 1, where the MFE and DTFN approaches notably outperformed other methodologies, ranking 1^{st} and 2^{nd} in the English segment of the official leaderboard, and 4^{th} and 13^{th} in the combined English and Spanish languages, respectively.

The remainder of this paper is structured as follows: Section 2 provides an overview of related work in sexism detection and ensemble learning strategies. Section 3 briefly describes the datasets. Section 4 details our methodologies, including the DTFN and MFE techniques. Section 5 details Experimental Assessment and Section 6 presents the results and analysis, followed by our conclusions in Section 7.

2. Related Work

Significant research has been conducted by researchers on the detection of hate speech, cyberbullying and offensive language. However, despite the growing interest on the topic, the literature on sexism detection is still limited.

In recent studies large language models (LLMs) and Transformer-based architecutres have been used for multi-modal detection of hate speech, sexism and offensive language from the text, images, memes, audio, and videos[8, 9, 10, 11, 12, 13, 14, 15]. The advent of transformer models, particularly BERT (Bidirectional Encoder Representations from Transformers) introduced by [16], enables a more sophisticated understanding of contextual relationships within text. Building upon BERT, [1] introduced RoBERTa (A Robustly Optimized BERT Approach), which fine-tuned the training process, and DeBERTa [2] (Decoding-enhanced BERT with Disentangled Attention), which incorporates a disentangled attention mechanism and enhanced decoding capabilities. Based on this pre-trained models, [17] fine-tuned deep learning models, such as CNN-BiLSTM and GPT-2, on the "MultiHate" dataset, achieving notable accuracy rates in sexism classification.

Moreover, the problem of sexism detection involves identifying harmful and biased language, often embedded within complex social contexts. Early approaches relied heavily on traditional machine learning techniques, such as SVMs and logistic regression, combined with manually crafted features Gaydhani et al. [18], Anistya and Setiawan [19]. However, these methods struggled with the subtleties of natural language and the contextual nature of sexism. Recent studies have leveraged the transformer models to address these challenges. In particular, [20] applied BERT to detect misogyny in social media. Similarly, Singh et al. [21] focused on the automatic detection of misogyny in multimodal online content by developing a large, annotated corpus of memes involving Hindi-English code-mixed language.

Ensemble approaches, which involves combining multiple models, have proven effective in various NLP tasks, and the rationale is that different models can capture different aspects of the data based

on their architecture, training objectives and data; thus, their combination can mitigate individual weaknesses. [22] provided a comprehensive overview of ensemble methods, emphasizing their potential to enhance robustness and accuracy. In the context of text classification, Stacked Generalization [23] and Bagging [24] are widely used ensemble techniques. More recent studies have focused on applying these methods to deep learning models. For example, [7] employed an ensemble of BERT-based models for sentiment analysis.

3. Datasets

In this study, we utilized the EXIST 2024 Tweets Dataset, specifically tailored for Task 1, which involves sexism identification in tweets. This dataset is comprehensive, containing over 10,000 labeled tweets balanced between English and Spanish. The tweets are annotated for binary classification, where the task is to determine whether a tweet contains sexist expressions or behaviors, categorized as "YES" or "NO."

The dataset is split into three parts: training, development, and test sets. For our experiments, we used the training and development sets with hard labels (gold standard) for model training and validation. The detailed distribution of the dataset is described in Table 1.

Table 1
Distribution of Data for Task 1: Sexism Detection in the EXIST 2024 Dataset

Language	Training Set	Development Set	Test Set
English	3,260	489	978
Spanish	3,460	549	1,098
Both Languages	6,920	1,038	2,076

4. Methodology

We proceed by introducing the two methodologies designed to address the classification of online sexism. First, we first present a simple yet effective method, named *Dual-Transformer Fusion Network* (DTFN) (see Section 4.1), based on the fusion of vector representations generated by two different neural language models; we subsequently present a more complex and effective approach, Multimodel Fusion Ensemble (MFE) (Section 4.2), based on ensemble learning of several LLMs and of the aforementioned DTFN.

4.1. Dual-Transformer Fusion Network (DTFN)

In our participation in the EXIST Task 1, we introduced a methodology named *Dual-Transformer Fusion Network* (DTFN), which integrates two Transformer models known for their effectiveness in online post analysis, namely RoBERTa-Large and DeBERTa-V3-Large [1, 2]. The DTFN methodology leverages the distinctive characteristics of each constituent model to enhance text classification: RoBERTa-Large, optimized for deep contextual understanding across longer text sequences [1], and DeBERTa-V3-Large, designed to model the inter-token relationships through its disentangled attention mechanism [2].

Based on these observations, we design DTFN as a hybrid architecture that first processes input text —typically extracted from social media or other online platforms— through both models in parallel. Each model then independently analyzes the text and outputs dense representation from their respective last hidden layers, potentially encoding complementary aspects of the text's semantic and contextual nuances.

Formally, let $\mathbf{x}$ denote the input text vector. Each Transformer model T (RoBERTa and DeBERTa) processes $\mathbf{x}$ independently and outputs a representation $\mathbf{h}_T$ from its final hidden layer:

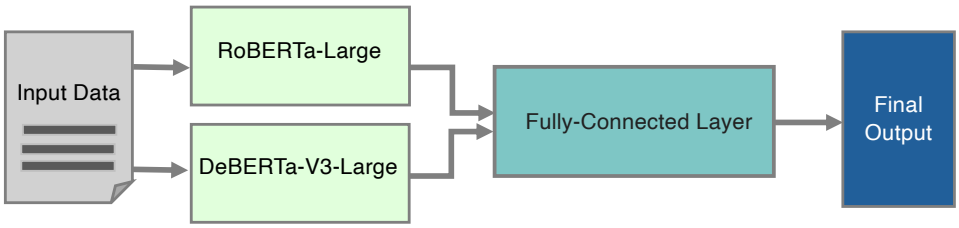

Figure 1: Architecture of the Concatenated Transformer Integration (DTFN) Technique. This diagram illustrates how outputs from the last layers of RoBERTa-Large and DeBERTa-V3-Large are concatenated and processed through a classifier(linear Layer) for enhanced sexism detection in multilingual contexts

$$\mathbf{h}_R = T_{\text{RoBERTa}}(\mathbf{x}), \quad \mathbf{h}_D = T_{\text{DeBERTa}}(\mathbf{x}) \tag{1}$$

These output vectors, $\mathbf{h}_R$ and $\mathbf{h}_D$, capture complementary linguistic features as determined by their distinct training paradigms and architectural innovations. Following the feature extraction, the outputs are concatenated to form a unified feature representation $\mathbf{h}$:

$$\mathbf{h} = [\mathbf{h}_R; \mathbf{h}_D] \tag{2}$$

This vector $\mathbf{h}$ is then passed through a fully connected linear layer L to produce the final class prediction $\hat{y}$. The linear layer acts as a classifier, integrating the diverse features into a unified prediction:

$$\hat{y} = \sigma(L(\mathbf{h})) \tag{3}$$

where σ denotes the sigmoid activation function, mapping the linear combination of features to a probability score indicating the final class predicted. The entire architecture is trained end-to-end with the objective of minimizing the binary cross-entropy loss $\mathcal{L}$:

$$\mathcal{L}(\hat{y}, y) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}) \tag{4}$$

The linear layer, as the whole architecture, is trained end-to-end on the specific task of sexism detection. Figure 1 illustrates the overall pipeline of the DTFN, highlighting the flow from the input text to the final classification.

4.2. Multimodel Fusion Ensemble (MFE)

Based on the promising results of our preliminary study combining two Transformer architectures, we devise a principled approach based on ensemble to dynamically combine multiple models with different architectures and training strategies. This approach, which we named Multimodel Fusion Ensemble (MFE), combines multiple models with distinct architectures and training strategies. Specifically, MFE integrates outputs from four different Transformer-based models — RoBERTa-Large[1], DeBERTa-V3-Large[2], Mistral-7b[3], and the previously introduced Dual-Transformer Network (DTFN) — using a majority voting mechanism.

Each model in the ensemble was selected for its unique capabilities in processing and understanding complex text structures, such as the dynamic masking strategy [1], the disentangled attention mechanism [2], the Grouped-Query and Sliding Window Attention [3]. Specifically, MFE integrates outputs from four different Transformer-based models - RoBERTa-Large, DeBERTa-V3-Large, Mistral-7b, and the previously introduced Dual-Transformer Network (DTFN) — using a majority voting mechanism. The individual models were first fine-tuned on the available dataset for sexism detection to optimize their

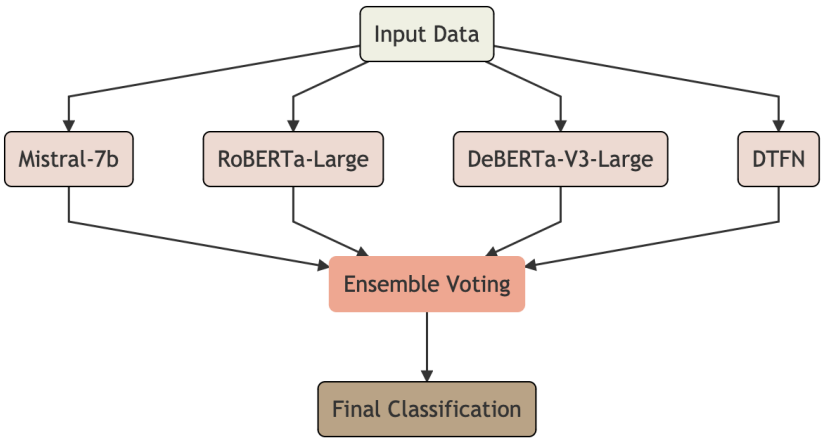

Figure 2: Architecture of the Multimodel Fusion Ensemble (MFE) Technique. This diagram shows the majority voting process involving RoBERTa-Large, DeBERTa-V3-Large, Mistral-7b, and the DTFN, optimized for enhanced sexism detection across multilingual contexts.

performance for the classification task. Subsequently, the ensemble was configured to employ a majority voting system to aggregate the predictions from each model.

More formally, in our ensemble method the classification decision for each instance is derived through a majority voting mechanism among the outputs of the constituent models. Let $C = \{c_1, c_2, \ldots, c_K\}$ represent the set of possible classes. For a given text instance x, each model m in the ensemble $M = \{m_1, m_2, \ldots, m_N\}$ predicts a class c_m. The ensemble prediction $\hat{c}$ is determined by:

$$\hat{c}(x) = \arg\max_{c \in C} \sum_{m=1}^{N} \mathbf{1}(c_m(x) = c) \tag{5}$$

where $\mathbf{1}$ is the indicator function that equals 1 if the condition is true and 0 otherwise. This simple approach counts the votes for each class from all models and selects the class with the highest count.

Majority Voting and Tie Handling

In scenarios where the voting results in a tie, particularly when the ensemble is evenly split across classes, a predefined rule is applied to resolve the ambiguity. Considering the sensitivity and potential consequences of misclassifying sexist content, our tie-breaking strategy defaults to the "Yes" (*Sexist*) prediction. This decision was based on the task's sensitivity and the potential social impact of under-detecting sexist content.

5. Experimental Assessment

As part of the EXIST 2024, we conducted a thorough experimental evaluation of the presented methodologies addressing, in particular, Task 1. In our experimental assessment, we initially evaluated the performance of individual models to establish a baseline. Then, we analysed the results yielded by the proposed Dual-Transformer Fusion Network (DTFN) and Multimodel Fusion Ensemble (MFE).

We proceed by first introducing the baselines, hyperparameters, and evaluation metrics adopted. We conclude by discussing the results on the EXIST dataset and the official leaderboard, along with quantitative analyses of the ensemble mechanism.

Baselines

In the following, we briefly describe the baselines evaluated:

- **RoBERTa-Large [1]:** An optimized version of BERT, whose model's size allows for a deeper understanding of language context, making it ideal for analyzing the intricacies in English and Spanish.
- **DeBERTa-V3-Large [2]:** It improves upon the BERT and RoBERTa designs by deciphering the dependency between words in a sentence, introducing a disentangled attention mechanism.
- **Mistral-7b [3]:** A large-scale model optimized for both performance and throughput and tailored for multilingual understanding. Its large-scale training on diverse datasets makes it particularly adept at handling the complexities of both English and Spanish.

We adopted their pre-trained versions, available through the HuggingFace library[1].

Parameter Settings

For each model used in our experiments, we identified a set of optimal hyperparameters through preliminary testing. The selected hyperparameters include the number of training epochs, learning rate (η), batch size, and weight decay (λ). Table 2 presents the optimal hyperparameters for each model.

Table 2
Best Hyperparameters per Model

Hyperparameter	RoBERTa-Large	DeBERTa-V3-Large	Mistral-7b	DTFN
Number of Epochs	30	30	10	30
Learning Rate	6×10^{-6}	6×10^{-6}	1×10^{-4}	6×10^{-6}
Batch Size	16	16	16	4
Weight Decay	5×10^{-3}	5×10^{-3}	5×10^{-3}	5×10^{-3}

Evaluation Metrics

In the evaluation of Task 1 for EXIST 2024, the official metrics used are ICM-Hard, ICM-Hard Norm, and F1. The ICM metric, proposed by [25], is based on information theory and measures the similarity between system classifications and gold standard labels. The organizers have also provided a normalized version, ICM-Hard Norm, to account for dataset imbalances, ensuring fair comparisons across different test conditions. For this shared task, higher values of the ICM and ICM-Hard Norm metrics indicate a stronger alignment between system outputs and the ground truth, with higher values considered better.

6. Results

6.1. Experimental Results on the Development set

The evaluation of the baseline models shows differences in performance across RoBERTa-Large, DeBERTa-V3-Large, Mistral-7b, and our proposed Dual-Transformer Fusion Network (DTFN), as reported in Table 3.

RoBERTa-Large achieved an F1 score of 0.864 and an ICM score of 0.592, demonstrating its robustness in handling the task. DeBERTa-V3-Large marginally outperformed RoBERTa-Large. Mistral-7b, on the other hand, yielded a slightly lower F1 score and the lowest ICM score among the baselines. This indicates that despite the higher number of parameters, Mistral-7b might not be as well-suited to the specific task of sexist identification compared to the other models evaluated.– Our proposed model, the Dual-Transformer Fusion Network (DTFN), slightly surpassed the other baseline models with an F1 score of 0.868 and showed a significant improvement with an ICM score of 0.606. The higher performance of DTFN highlights the efficacy of our dual-transformer architecture in improving classification accuracy.

[1] https://huggingface.co/

Table 3
Performance Metrics of Baseline and Ensemble Model Combinations on the Development Set

ID	Models	F1 Score	ICM Score
	Baseline Models		
1	RoBERTa-Large	0.864	0.592
2	DeBERTa-V3-Large	0.866	0.598
3	Mistral-7b	0.859	0.577
	Dual-Transformer Fusion Architecture		
4	*Dual-Transformer Fusion Network (DTFN)* - Ours	**0.868**	**0.606**
	Ensemble Combinations		
5	RoBERTa-Large, DeBERTa-V3-Large, Mistral-7b	0.8811	0.6438
6	RoBERTa-Large, DeBERTa-V3-Large, *DTFN*	0.8811	0.6439
7	RoBERTa-Large, Mistral-7b, *DTFN*	0.8832	0.6507
8	DeBERTa-V3-Large, Mistral-7b, *DTFN*	0.8747	0.6248
9	*Multimodel Fusion Ensemble (MFE)* - Ours RoBERTa-Large, DeBERTa-V3-Large, Mistral-7b, *DTFN*	**0.8841**	**0.6548**

Experimental Analysis of the Majority Voting

To systematically understand the voting results, we analysed the number of times each combination of models agreed on the sexist (Yes) and non-sexist (No) classes. This is done to determine whether there is a dominant model (or combination) in the ensemble. The combinations and their respective agreement counts are detailed in Table 4. The table provides detailed insights into the agreement and non-conformity of various ensemble model combinations on the development set.

Among the combinations, the ensemble of RoBERTa-Large, DeBERTa-V3-Large, Mistral-7b, and DTFN shows the highest majority agreement. Conversely, the combination of RoBERTa-Large, DeBERTa-V3-Large, and Mistral-7b without DTFN showed the lowest majority agreement. This indicates that the addition of DTFN significantly boosts the ensemble's agreement, particularly in identifying sexist content. Out of all instances, we observed a total of 58 ties where the aforementioned tie-breaking rule was applied.

Additionally, we reported the isolation frequency, i.e., how often a single model's prediction differed from the majority vote within the ensemble, reflecting the model's conformity with others. RoBERTa-Large had the highest isolation frequency, which could be due to the fact that it takes context into account. DeBERTa-V3-Large showed a lower isolation frequency, while Mistral-7b frequently disagreed

Table 4
Majority Voting Agreement and Non-conformity Metrics in Multimodel Fusion Ensemble (MFE) on Development Set

Ensemble Combination	Sexist ('Yes')	Non-Sexist ('No')
RoBERTa-Large, DeBERTa-V3-Large, and Mistral-7b	11	11
RoBERTa-Large, DeBERTa-V3-Large, and DTFN	16	27
RoBERTa-Large, Mistral-7b, and DTFN	19	7
DeBERTa-V3-Large, Mistral-7b, and DTFN	25	13
RoBERTa-Large, DeBERTa-V3-Large, Mistral-7b, and DTFN	386	361
Isolation Frequency		
RoBERTa-Large	25	13
DeBERTa-V3-Large	19	7
Mistral-7b	16	27
DTFN	11	11

Table 5
Official Results on the Test Set for Task 1 (Sexism Detection) Using Dual-Transformer Fusion Network (DTFN) and Multimodel Fusion Ensemble (MFE)

Evaluation	Language	Approach	ICM-Hard	ICM-Hard Norm	F1	Rank
Hard-Hard	English	MFE	**0.6178**	**0.8153**	**0.7610**	**1st/68**
Hard-Hard	English	DTFN	0.5953	0.8038	0.7491	**2nd/68**
Hard-Hard	Spanish	MFE	**0.5497**	**0.7748**	**0.7898**	12^{th}/66
Hard-Hard	Spanish	DTFN	0.4903	0.7452	0.7710	25^{th}/66
Hard-Hard	Both	MFE	**0.5883**	**0.7956**	**0.7775**	4^{th}/70
Hard-Hard	Both	DTFN	0.5447	0.7738	0.7614	13^{th}/70

with the ensemble on non-sexist classifications. DTFN had the lowest isolation frequency, suggesting it is the most conforming model within the ensemble, underscoring the DTFN's role in enhancing ensemble cohesion.

6.2. Results of the Official Leaderboard

In this section, we present the results of our participation in Task 1 of the EXIST 2024 challenge, where our team, EquityExplorers, submitted two runs: EquityExplorer-1 using the DTFN technique and EquityExplorer-2 employing the Multimodel Fusion Ensemble (MFE) approach.

The MFE and the DTFN demonstrated notable performance by ranking 1^{st} and 2^{nd} for the Task 1 in the English segment, respectively. The effectiveness of the MFE, evidenced by its ICM-Hard score of 0.6178 and ICM-Hard Norm of 0.8153, coupled with an F1 score of 0.7610, demonstrated its robust capability to discern nuances of sexist content in English tweets effectively. This ensemble approach, by combining different strategies and model outputs, has proven to be particularly effective in improving accuracy and reliability over individual models, including the Dual-Transformer Fusion Network.

The patterns observed in the Spanish and Both (combining results from English and Spanish) evaluations align with these findings. Although the performance gap in the Spanish evaluation is wider, it highlights the robustness of MFE in a different linguistic environment. Notably, DTFN ranks 25^{th} out of 66 in Spanish, suggesting that while it is effective, it might not fully adapt to different languages as efficiently as MFE. The aggregated results for both languages demonstrate the consistent advantage of using an ensemble approach, with MFE achieving the 4^{th} rank out of 70, compared to the 13^{th} rank for DTFN. In conclusion, the official leaderboard results validate the proposed approach and highlight the significant improvements achieved through the Multimodel Fusion Ensemble (MFE). The consistent outperformance of MFE across various metrics and datasets underscores the potential of ensemble methods involving neural language models.

7. Conclusion

In this work, we introduced the Dual-Transformer Fusion Network (DTFN) and the Multimodel Fusion Ensemble (MFE) for identifying sexist content across multiple languages within the context of the EXIST 2024 competition. A thorough evaluation on the development and test sets highlighted the superior performance of the MFE, particularly ranking highly in both the English (1^{st}) and combined language (4^{th}) categories. This performance, paired with a comparative analysis against baseline models, allowed for a detailed assessment of the relative improvements offered by the DTFN and MFE approaches. It demonstrated the benefits of integrating diverse transformer models into an ensemble framework to leverage the characteristics of each neural language model, thereby achieving higher accuracy and reliability in detecting complex linguistic patterns associated with sexism.

References

[1] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.

[2] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, 2021. arXiv:2006.03654.

[3] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. arXiv:2310.06825.

[4] H. Yan, L. Gui, G. Pergola, Y. He, Position bias mitigation: A knowledge-aware graph model for emotion cause extraction, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), 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), Association for Computational Linguistics, Online, 2021, pp. 3364–3375.

[5] 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.

[6] 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.

[7] A. C. Mazari, N. Boudoukhani, A. Djeffal, Bert-based ensemble learning for multi-aspect hate speech detection, Cluster Computing 27 (2024) 325–339.

[8] D. Kikkisetti, R. U. Mustafa, W. Melillo, R. Corizzo, Z. Boukouvalas, J. Gill, N. Japkowicz, Using llms to discover emerging coded antisemitic hate-speech in extremist social media, 2024. arXiv:2401.10841.

[9] L. Zhu, G. Pergola, L. Gui, D. Zhou, Y. He, Topic-driven and knowledge-aware transformer for dialogue emotion detection, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), 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), Association for Computational Linguistics, Online, 2021, pp. 1571–1582.

[10] G. Pergola, L. Gui, Y. He, A disentangled adversarial neural topic model for separating opinions from plots in user reviews, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 2870–2883.

[11] R. Wolfe, Y. Yang, B. Howe, A. Caliskan, Contrastive language-vision ai models pretrained on web-scraped multimodal data exhibit sexual objectification bias, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 1174–1185. URL: https://doi.org/10.1145/3593013.3594072. doi:10.1145/3593013.3594072.

[12] G. Pergola, L. Gui, Y. He, TDAM: A topic-dependent attention model for sentiment analysis, Information Processing & Management 56 (2019) 102084.

[13] J. Lu, X. Tan, G. Pergola, L. Gui, Y. He, Event-centric question answering via contrastive learning and invertible event transformation, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 2377–2389.

[14] J. Lu, J. Li, B. Wallace, Y. He, G. Pergola, NapSS: Paragraph-level medical text simplification via narrative prompting and sentence-matching summarization, in: A. Vlachos, I. Augenstein

(Eds.), Findings of the Association for Computational Linguistics: EACL 2023, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 1079–1091.
[15] A. Irfan, D. Azeem, S. Narejo, N. Kumar, Multi-modal hate speech recognition through machine learning, in: 2024 IEEE 1st Karachi Section Humanitarian Technology Conference (KHI-HTC), 2024, pp. 1–6. doi:10.1109/KHI-HTC60760.2024.10482031.
[16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.
[17] A. Vetagiri, P. Pakray, A. Das, A deep dive into automated sexism detection using fine-tuned deep learning and large language models, Available at SSRN 4791798 (2024).
[18] A. Gaydhani, V. Doma, S. Kendre, L. Bhagwat, Detecting hate speech and offensive language on twitter using machine learning: An n-gram and {TFIDF} based approach, CoRR abs/1809.08651 (2018). URL: http://arxiv.org/abs/1809.08651. arXiv:1809.08651.
[19] F. Anistya, E. B. Setiawan, Hate speech detection on twitter in indonesia with feature expansion using glove, Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi) 5 (2021) 1044 – 1051. URL: http://www.jurnal.iaii.or.id/index.php/RESTI/article/view/3521. doi:10.29207/resti.v5i6.3521.
[20] A. Rahali, M. A. Akhloufi, A.-M. Therien-Daniel, E. Brassard-Gourdeau, Automatic misogyny detection in social media platforms using attention-based bidirectional-lstm*, in: 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2021, pp. 2706–2711. doi:10.1109/SMC52423.2021.9659158.
[21] A. Singh, D. Sharma, V. K. Singh, Mimic: Misogyny identification in multimodal internet content in hindi-english code-mixed language, ACM Trans. Asian Low-Resour. Lang. Inf. Process. (2024). URL: https://doi.org/10.1145/3656169. doi:10.1145/3656169, just Accepted.
[22] T. G. Dietterich, Ensemble methods in machine learning, in: International workshop on multiple classifier systems, Springer, 2000, pp. 1–15.
[23] D. H. Wolpert, Stacked generalization, Neural Networks 5 (1992) 241–259.
[24] L. Breiman, Bagging predictors, Machine Learning 24 (1996) 123–140. URL: https://api.semanticscholar.org/CorpusID:47328136.
[25] E. Amigo, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5809–5819.