FIGHTING FAKE-NEWS VIA ADAPTIVE MIXTURE OF TRANSFORMERS

Alexandru Petrescu[1], Patricia Vasile[2]

The increase of online political propaganda and misinformation presents a significant challenge to the integrity of public discourse. This paper proposes an adaptive Mixture of Transformers (MOT) framework for the detection of Misinformation and extremist content across multiple domains. The methodology integrates advanced Transformer-based models, including ALBERT, RoBERTa, and DeBERTa variants, within a unified pipeline that encompasses data aggregation, preprocessing, model training, and evaluation. Experiments leverage diverse, publicly available datasets and employ a standardized multi-class labeling scheme to ensure robust, generalizable results. The evaluation demonstrates that the MOT ensemble consistently outperforms individual models in terms of F1-score and recall, particularly when datasets are combined. The approach is further validated against graph neural network baselines, highlighting its scalability and suitability for automated content moderation systems.

Keywords: Misinformation detection, propaganda detection, Mixture of Transformers, ensemble learning

1. Introduction

Online misinformation has become an increasingly serious threat due to the growing popularity of social networks. Misinformation refers to false, inaccurate, or misleading information that is spread, regardless of an intent to deceive. Unlike disinformation, which is deliberately created and disseminated to mislead or manipulate, misinformation is often shared by people who believe it to be true. This growing challenge has led the scientific community to focus on developing automated detection systems [9]. Thanks to recent advances in natural language processing, especially Transformer-based models, modern AI systems now significantly outperform traditional methods in identifying propaganda and disinformation at scale.

Detecting and managing misinformation remains a complex challenge, as a single text can simultaneously employ multiple manipulative techniques, making

[1] PhD student Eng., National University of Science and Technology Politehnica University Bucharest, Romania, and Academy of Romanian Scientists, Romania, corresponding author, e-mail: alex.petrescu@upb.ro

[2] Eng., National University of Science and Technology Politehnica University Bucharest, Romania, e-mail: patriciavasile2700@gmail.com

conventional filtering methods insufficient. Automated and scalable solutions capable of analyzing large volumes of online content in real-time are therefore essential to effectively distinguish legitimate discourse from intentional misinformation or hate speech [38].

Our objectives for this article are:

- **Misinformation Detection:** Identification of clues and patterns associated with misinformation, for example radical or extremist political propaganda .
- **Machine Learning Algorithms:** Integrating advanced Transformer models (e.g. BERT) to enhance detection accuracy [39].
- **Advanced NLP Pipelines:** Extracting, cleaning, standardizing, and efficiently representing textual features for classification.
- **Automated Mitigation Strategies:** Implementing automatic moderation and filtering mechanisms to reduce misinformation impact.
- **Scalability:** Ensuring real-time processing of large data volumes and providing visualization tools to monitor system performance.
- **Methodology Focus:** Primarily centered on textual content analysis [37], aiming to protect users and foster a responsible digital environment.

In this article, we begin with a review of related work in misinformation detection, providing context and highlighting previous approaches in the field, chapter 2. The methodology, chapter 3, follows, detailing the pipeline for detecting misinformation content. This is further broken down into a discussion of the datasets used for misinformation detection, each described in its own subsection. The experiments, chapter 4, presents the evaluation framework and results, with additional subsections dedicated to the components of the Mixture of Transformers (MOT) architecture. The article concludes with a summary, chapter 5, where we discuss findings and implications.

2. Related Work in Misinformation Detection

In recent years, researchers have explored different approaches to detecting and mitigating the spread of misinformation. The most common strategy integrates transformer-based embeddings with deep learning architectures to create models for classifying textual information [15, 29, 30, 32, 22, 7, 24, 25]. Another research direction emphasizes enhancing contextual understanding by incorporating metadata, such as social dynamics [33] or the propagation patterns of content across networks [35]. Moreover, some research directions propose the use of network immunization strategies to stop the online dissemination of harmful content [28, 20, 31, 6, 23]. Complementing these efforts, end-to-end real-time systems have also been developed to continuously monitor and analyze social media platforms in real-time, aiming to detect and stop the spread of harmful content [5, 34].

Current research in the field of misinformation has evolved significantly beyond simple descriptive statistics or vocabulary visualization. Today, experts use semantic representations and complex architectures that can identify stylistic, syntactic, and pragmatic features specific to manipulative discourse.

3. Methodology

The methodology adopted in this work is organized as a pipeline, encompassing several critical stages to ensure robust and reliable detection of misinformation content. The initial phase involves data collection and preprocessing, which includes aggregating textual data from diverse sources such as Twitter posts and news articles. This stage further incorporates standard NLP techniques, including tokenization, normalization, and the removal of irrelevant terms, to prepare the data for subsequent analysis.

Following preprocessing, the next stage focuses on the training and fine-tuning of Transformer-based classification models. These models are trained using labeled datasets that contain unbiased content, enabling the system to generalize effectively across various forms of political discourse.

The performance of the trained models is rigorously evaluated using a suite of standard metrics, including precision, recall, and F1-Score. This evaluation facilitates the identification of the models' strengths and weaknesses.

Central to our approach is the proposed Mixture of Transformers (MOT) architecture, initially introduced in [22]. The MOT framework enables the evaluation of individual Transformer models as well as their performance in an ensemble configuration, thereby leveraging the complementary strengths of multiple architectures to enhance overall detection accuracy.

3.1. Data sets for misinformation detection

This chapter provides an overview of several publicly available datasets that are frequently utilized in academic and applied research aimed at the detection of misinformation. These datasets serve as valuable resources for developing and benchmarking algorithms designed to identify and mitigate the spread of misleading or harmful information online.

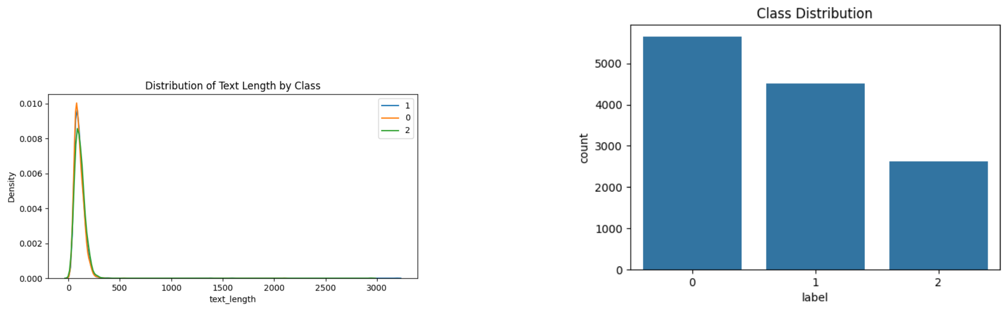
To facilitate a consistent and meaningful comparison across the selected datasets, we propose a unified multi-class classification. This framework categorizes each data instance into one of the following three distinct classes:

- **0 = True:** This class includes content that is verified as factual, authentic, or otherwise trustworthy (e.g., Real, Non-Rumor, etc.).
- **1 = Misinformation:** This category encompasses content identified as false, deceptive, manipulative, or intended to mislead.
- **2 = Cannot Decide:** This class is reserved for cases where the veracity of the content cannot be confidently determined due to insufficient information, ambiguity, or lack of consensus among annotators.

By adopting this standardized classification scheme, we aim to harmonize the labeling conventions across diverse datasets, thereby enabling more robust and generalizable evaluation of detection methods and fostering greater reproducibility in future research.

3.1.1. *LIAR dataset.* The LIAR dataset [41] represents a sophisticated and widely used resource for the study of misinformation and political discourse. It consists of a large collection of short statements sourced from PolitiFact.com, each meticulously annotated with one of six fine-grained veracity labels. This nuanced labeling scheme facilitates in-depth linguistic and stylistic analyses, enabling researchers to explore the subtle characteristics of propaganda, ideological framing, and contextual distortions within political communication [26].

As illustrated in Figure 1, the majority of the dataset's content is at the tweet level, reflecting the concise nature of the statements analyzed. Furthermore, the distribution of instances among the `0` (True) and `1` (Misinformation) classes is relatively balanced, which is advantageous for training and evaluating classification models. However, it is important to note that the `2` (Cannot Decide) class exhibits a significant imbalance, a trend that is expected to persist across experimental setups. This imbalance should be carefully considered when interpreting model performance, as it may impact both the training process and the generalizability of the results.



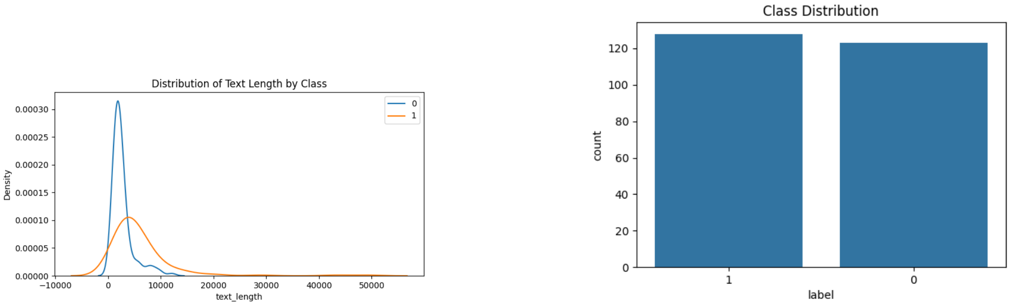
(A) Length - LIAR

(B) Class - LIAR

FIG. 1. LIAR text length and class distributions

3.1.2. *Horne 2017 Dataset.* The Horne 2017 dataset [12] is a valuable resource for research in misinformation and media studies. It comprises news articles systematically labeled as real, fake, or satirical, with sources drawn from a diverse range of mainstream news outlets, misinformation sites, and satire platforms. The dataset was specifically curated to facilitate the examination of linguistic and structural differences among real, fake, and satirical news, making it particularly well-suited for studies focused on stylistic features such as readability, headline complexity, and lexical usage patterns.

As illustrated in Figure 2, the dataset is relatively modest in size compared to other resources, and its class distribution is balanced between the `0` (True) and `1` (Misinformation) categories. Notably, the `2` (Cannot Decide) class is absent from this dataset, which should be taken into account when comparing results across datasets with different labeling schemes.



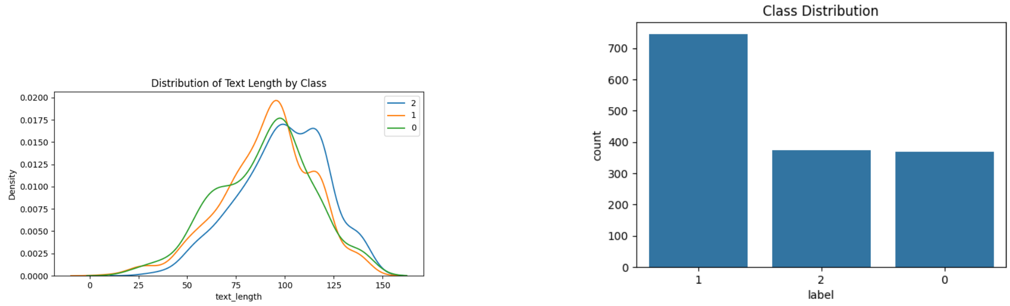
(A) Length - HORNE2017

(B) Class - HORNE2017

Fig. 2. HORNE2017 text length and class distributions

3.1.3. *Twitter15 & Twitter16 Dataset.* The Twitter15 and Twitter16 datasets are influential resources in the study of rumor detection and stance classification on Twitter. Each dataset comprises annotated tweets, accompanied by their complete reply threads, allowing for the analysis of conversational context and information propagation. Specifically, Twitter15 contains approximately 1,500 distinct events, while Twitter16 includes around 800 events, with each event representing a rumor or news story that sparked discussion.

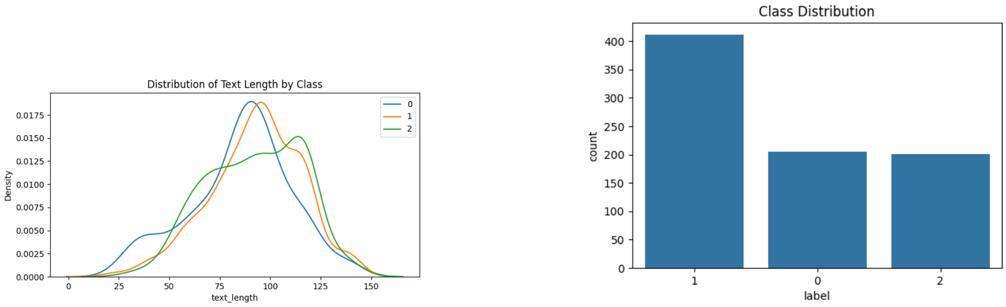
As depicted in Figures 3 and 4, both datasets exhibit a balanced distribution across the available classes, including the 2 (Cannot Decide/Unverified) category. Notably, these datasets present an interesting scenario in which the 0 (True) and 2 (Unverified) classes are approximately balanced, while the 1 (Misinformation/False) class contains as many examples as the other two classes combined. This unique distribution poses interesting challenges and opportunities for model development, particularly in handling class imbalance and capturing the subtleties of misinformation in social media contexts.



(A) Length - Twitter15

(B) Class - Twitter15

Fig. 3. Twitter15 text length and class distributions



(A) Length - Twitter16 (B) Class - Twitter16

Fig. 4. Twitter16 text length and class distributions

4. Experiments

The proposed framework first assesses individual models and the MOT architecture on each dataset separately, followed by evaluation on a combined dataset encompassing all sources. Evaluation was performed using an 80/20 stratified train/test split by class. Reported metrics include accuracy, macro-averaged precision, recall, and F1-score. The pipeline's main goal is to maximize F1, with an early stop of 3 and a maximum of 50, that is never reached in our experiments, as the models reach saturation fast on small datasets.

4.1. MOT Components

For our MOT architecture, we utilize the following models:

- **microsoft/Multilingual-MiniLM-L12-H384** [42]: A distilled multilingual Transformer with 21 million parameters, designed for cross-lingual natural language inference and question answering.
- **distilbert/distilbert-base-uncased-finetuned-sst-2-english** [36]: A DistilBERT model fine-tuned on the SST-2 sentiment analysis dataset.
- **facebook/roberta-hate-speech-dynabench-r4-target** [40]: A RoBERTa-base model adversarially fine-tuned for hate speech detection in English.
- **KoalaAI/Text-Moderation** [1]: A DeBERTa-v3 based moderation model, trained on data aligned with OpenAI's content policy.
- **jy46604790/Fake-News-Bert-Detect** [2]: A RoBERTa-base model for fake news detection, trained on a large dataset of fake and real news.
- **XSY/albert-base-v2-fakenews-discriminator** [3]: An ALBERT-base model trained exclusively on news headlines.

[1] KoalaAI. *KoalaAI/Text-Moderation: DeBERTa-v3 based text moderation model.* Hugging Face. 2024. Available at: `https://huggingface.co/KoalaAI/Text-Moderation`

[2] jy46604790. *Fake-News-Bert-Detect: RoBERTa-base model for fake news detection.* Hugging Face: `https://huggingface.co/jy46604790/Fake-News-Bert-Detect`

[3] XSY. *albert-base-v2-fakenews-discriminator.* Hugging Face: `https://huggingface.co/XSY/albert-base-v2-fakenews-discriminator`

- **vikram71198/distilroberta-base-finetuned-fake-news-detection** [4]: A DistilRoBERTa model trained on over 40,000 examples from various Kaggle fake news datasets.

The baseline proposed in this methodology is ALBERT [16], chosen for its efficiency and reduced risk of overfitting on smaller datasets, but, as can be seen in Figure 5, it averages the highest training time. An important thing to consider is that the MOT adds no overhead for the proposed datasets in terms of training time, which is why it is not represented in the Figure 5.

4.2. Results

In all our experiments, we notice that MOT behaves better than the average of its components for all the metrics and all datasets. Interestingly, our experiments show that its average F1-Score [27] is not the best when it comes to individual datasets, as depicted in Figure 6a, but when we combine the datasets, it becomes best, as depicted in Figure 6b. The same phenomenon can be seen for Recall, Figure 7, but not for precision due to the baseline model, as can be seen in Figure 8, but still above average.

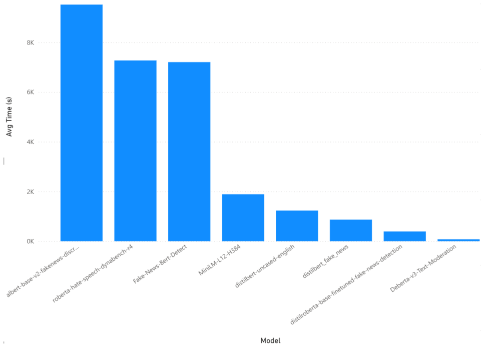
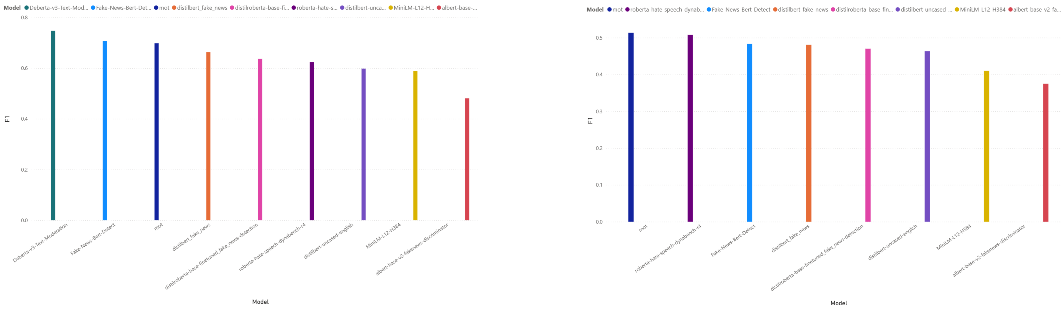


Fig. 5. Average training time (in seconds) for each model



(A) Avg F1 over individual datasets

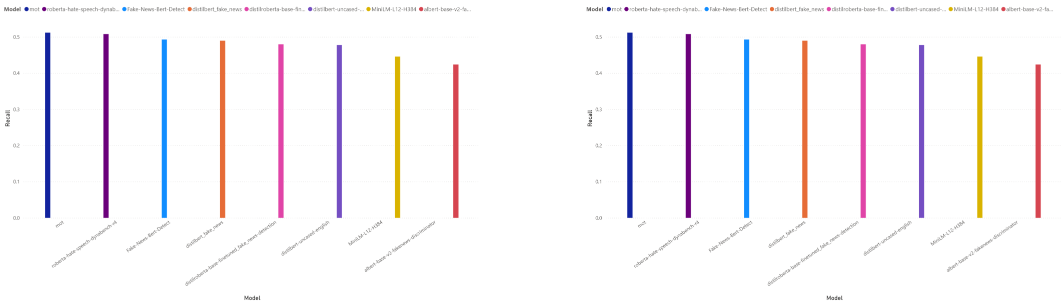
(B) Avg F1 over the combined dataset

Fig. 6. F1 Score

The difference in individual vs due to the liar dataset, where we have:

- **F1**: avg 0.41 / max 0.44 (MOT)
- **Precision**: avg 0.47 / max 0.7 (roberta-hate-speech-dynabench-r4-target)

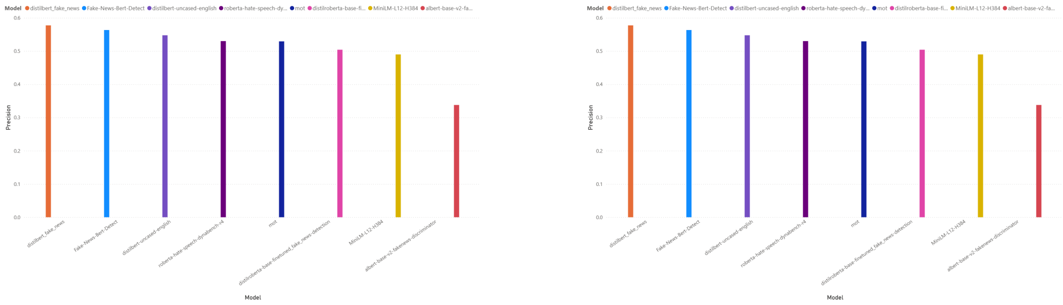
[4] vikram71198. *distilroberta-base-finetuned-fake-news-detection: DistilRoBERTa model for fake news detection.* Hugging Face. 2022. Available at: `https://huggingface.co/vikram71198/distilroberta-base-finetuned-fake-news-detection`



(A) Average Recall over individual datasets

(B) Average Recall over the combined dataset

Fig. 7. Recall Score



(A) Average Precision over individual datasets

(B) Average Precision over the combined dataset

Fig. 8. Precision Score

- **Recal**: avg 0.44 / max 0.46 (MOT)

This behavior can be expected due to the nature of the content used in the pipelines, the headlines, as they fit in the area of the used models. The full text is also aivalbe, but the current setup can not be used for this, another one should be employed and would change the nature of the experiments.

Another consideration for this particular experiment is the imbalance of the classes, to be specific the "Cannot Decide" class, here MOT manages to score the highest metrics for Recal and F1, as previously mentioned, but they are both under 0.1 where the average is under 0.05.

We have decided to keep this experiment as it provides powerful insights and gives us future research opportunities.

5. **Conclusions**

This work demonstrates that an adaptive Mixture of Transformers (MOT) framework provides a robust and scalable solution for the detection of Misinformation and extremist political content. By integrating multiple Transformer-based models and leveraging diverse, publicly available datasets, the proposed system achieves improved generalization and higher detection accuracy compared to individual models. Experimental results show that the MOT ensemble consistently outperforms its components, especially when evaluated on combined datasets, with notable gains in F1-score and recall. The approach also maintains efficiency, introduces minimal training overhead, and proves suitable for real-world deployment in automated content moderation systems. Future work will explore the integration of multi-modal signals and advanced graph-based models to further enhance detection capabilities and adapt to evolving patterns of online misinformation.

For general future directions we plan to use different strategies in the MOT architecture, to better leverage the strengths and weaknesses of each component, as well as emphasize to the specific models for the given task.

For the interesting aspect that the experiments with the LIAR dataset we plan to expand on the analysis of length vs nature, as we have observed that while the headline of the article makes sense as of text length and nature of content (misinformation) considerations, current techniques yield unsatisfactory results. We can top stating that the nature of the content is important, as we had a multitude of topics for those headlines.

Acknowledgments

The research presented in this paper was supported in part by 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).

REFERENCES

[1] *Malak Abdullah, Dia Abujaber, Ahmed Al-Qarqaz, Rob Abbott, Mirsad Hadzikadic.* Combating propaganda texts using transfer learning, *IAES International Journal of Artificial Intelligence*, 12:956, 2023.

[2] *Hadeer Ahmed, Issa Traore, Sherif Saad.* Detecting opinion spams and fake news using text classification, *Security and Privacy*, 1(1):e9, 2018. DOI: `https://doi.org/10.1002/spy2.9.`

[3] *Abubakar Aliero, Sulaimon Bashir, Hamzat Aliyu, Amina Tafida, Bashar Kangiwa, Nasiru Dankolo.* Systematic review on text normalization techniques and its approach to non-standard words, *International Journal of Computer Applications*, 185:975–8887, 2023.

[4] *Hani Al-Omari, Malak Abdullah, Ola AlTiti, Samira Shaikh.* JUSTDeep at NLP4IF 2019 task 1: Propaganda detection using ensemble deep learning models, *Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, 113–118, 2019.

[5] *Elena-Simona Apostol, Ciprian-Octavian Truică, Adrian Paschke.* ContCommRTD: A Distributed Content-Based Misinformation-Aware Community Detection System for Real-Time Disaster Reporting, *IEEE Transactions on Knowledge and Data Engineering*, 36(11):5811–5822, 2024. DOI: `https://doi.org/10.1109/tkde.2024.3417232`.

[6] *Elena-Simona Apostol, Özgur Coban, Ciprian-Octavian Truică.* CONTAIN: A community-based algorithm for network immunization, *Engineering Science and Technology, an International Journal*, 55:1–10(101728), 2024. DOI: `https://doi.org/https://doi.org/10.1016/j.jestch.2024.101728`.

[7] *Maria-Diana Cotelin, Elena-Simona Apostol, Ciprian-Octavian Truică.* NetGuardAI at EXIST2025: Sexism Detection using mDeBERTa, *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025)*, 2025.

[8] *Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova.* BERT: Pre-training of deep bidirectional transformers for language understanding, *Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–4186, 2019.

[9] *Zhida Feng, Jiji Tang, Jiaxiang Liu, Weichong Yin, Shikun Feng, Yu Sun, Li Chen.* Alpha at SemEval-2021 Task 6: Transformer Based Propaganda Classification. *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, 99–104, 2021.

[10] *Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, Noah A. Smith.* Don't stop pretraining: Adapt language models to domains and tasks, *Annual Meeting of the Association for Computational Linguistics*, 8342–8360, 2020.

[11] *Pengcheng He, Xiaodong Liu, Jianfeng Gao, Weizhu Chen.* DeBERTa: Decoding-enhanced BERT with disentangled attention, *International Conference on Learning Representations (ICLR 2021)*, 2021.

[12] *Benjamin D. Horne, Sibel Adalı.* This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News, *NECO*, 2017.

[13] *Jeremy Howard, Sebastian Ruder.* Universal language model fine-tuning for text classification, *Annual Meeting of the Association for Computational Linguistics*, 328–339, 2018.

[14] *Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen.* LoRA: Low-Rank Adaptation of Large Language Models, *International Conference on Learning Representations (ICLR 2022)*, 2022.

[15] *Vlad-Iulian Ilie, Ciprian-Octavian Truică, Elena-Simona Apostol, Adrian Paschke.* Context-Aware Misinformation Detection: A Benchmark of Deep Learning Architectures Using Word Embeddings, *IEEE Access*, 9:162122–162146, 2021. DOI: `https://doi.org/10.1109/ACCESS.2021.3132502`.

[16] *Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut.* ALBERT: A lite BERT for self-supervised learning of language representations, *International Conference on Learning Representations*, 2020.

[17] *Brian Lester, Rami Al-Rfou, Noah Constant.* The Power of Scale for Parameter-Efficient Prompt Tuning, *Conference on Empirical Methods in Natural Language Processing*, 3045–3059, 2021.

[18] *Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov.* RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv preprint arXiv:1907.11692, 2019.

[19] *Alexandru Petrescu, Ciprian-Octavian Truică, Elena-Simona Apostol.* Sentiment Analysis of Events in Social Media, *International Conference on Intelligent Computer Communication and Processing (ICCP 2019)*, 143–149, 2019. DOI: `https://doi.org/10.1109/iccp48234.2019.8959677`.

[20] *Alexandru Petrescu, Ciprian-Octavian Truică, Elena-Simona Apostol, Panagiotis Karras.* Sparse Shield: Social Network Immunization vs. Harmful Speech, ACM International Conference on Information and Knowledge Management (CIKM2021), 1426–1436, 2021. DOI: `https://doi.org/10.1145/3459637.3482481`.

[21] *Alexandru Petrescu.* Leveraging MiniLMv2 Pipelines for EXIST2023, *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023)*, 1037–1043, 2023.

[22] *Alexandru Petrescu, Ciprian-Octavian Truică, Elena-Simona Apostol.* Language-based Mixture of Transformers for EXIST2024, *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*, 1157–1164, 2024.

[23] *Alexandru Petrescu, Ciprian-Octavian Truică, Elena-Simona Apostol, Adrian Paschke.* EDSA-Ensemble: an Event Detection Sentiment Analysis Ensemble Architecture, *IEEE Transactions on Affective Computing*, 16(2):555-572, 2025. DOI: `https://doi.org/10.1109/TAFFC.2024.3434355`.

[24] *Alexandru Petrescu, Ciprian-Octavian Truică, Elena-Simona Apostol.* Language-based Mixture of Transformers for Sexism Identification in Social Networks, *Conference and Labs of the Evaluation Forum (CLEF 2025)*, 2025.

[25] *Alexandru Petrescu, Elena-Simona Apostol, Ciprian-Octavian Truică.* Awakened at EXIST2025: Adaptive Mixture of Transformers, *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025)*, 2025.

[26] *Yu Qiao, Daniel Wiechmann, Elma Kerz.* A language-based approach to fake news detection through interpretable features and BRNN, *International Workshop on Rumours and Deception in Social Media*, 14–31, 2020.

[27] *Ciprian-Octavian Truică, Cătălin Adrian Leordeanu.* Classification of an imbalanced data set using decision tree algorithms, *University Politechnica of Bucharest Scientific Bulletin - Series C Electrical Engineering and Computer Science*, 79:69–84, 2017.

[28] *Ciprian-Octavian Truică, Elena-Simona Apostol, Teodor Ștefu, Panagiotis Karras.* A Deep Learning Architecture for Audience Interest Prediction of News Topic on Social Media, *International Conference on Extending Database Technology (EDBT2021)*, 588–599, 2021. DOI: `https://doi.org/10.5441/002/EDBT.2021.69`.

[29] *Ciprian-Octavian Truică, Elena-Simona Apostol.* MisRoBÆRTa: Transformers versus Misinformation, *Mathematics*, 10:1–25(569), 2022. DOI: `https://doi.org/10.3390/math10040569`.

[30] *Ciprian-Octavian Truică, Elena-Simona Apostol, Adrian Paschke.* Awakened at CheckThat! 2022: Fake News Detection using BiLSTM and sentence transformer, *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2022)*, 749–757, 2022.

[31] *Ciprian-Octavian Truică, Elena-Simona Apostol, Radu-Cătălin Nicolescu, Panagiotis Karras.* MCWDST: A Minimum-Cost Weighted Directed Spanning Tree Algorithm for Real-Time Fake News Mitigation in Social Media, *IEEE Access*, 11:125861–125873, 2023. DOI: `https://doi.org/10.1109/ACCESS.2023.3331220`.

[32] *Ciprian-Octavian Truică, Elena-Simona Apostol.* It's all in the Embedding! Fake News Detection using Document Embeddings, *Mathematics*, 11:1–29(508), 2023. DOI: `https://doi.org/10.3390/math11030508`.

[33] *Ciprian-Octavian Truică, Elena-Simona Apostol, Panagiotis Karras.* DANES: Deep Neural Network Ensemble Architecture for Social and Textual Context-aware Fake News Detection, *Knowledge-Based Systems*, 294:1–13(111715), 2024. DOI: `https://doi.org/https://doi.org/10.1016/j.knosys.2024.111715`.

[34] *Ciprian-Octavian Truică, Ana-Teodora Constantinescu, Elena-Simona Apostol.* StopHC: A Harmful Content Detection and Mitigation Architecture for Social Media Platforms, *IEEE International Conference on Intelligent Computer Communication and Processing (ICCP 2024)*, 1–5, 2024. DOI: `https://doi.org/10.1109/ICCP63557.2024.10793051`.

[35] *Ciprian-Octavian Truică, Elena-Simona Apostol, Marius Marogel, Adrian Paschke.* GETAE: Graph Information Enhanced Deep Neural NeTwork Ensemble ArchitecturE for fake news detection, *Expert Systems with Applications*, 275:1–14(126984), 2025. DOI: `https://doi.org/10.1016/j.eswa.2025.126984`.

[36] *Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf.* DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108, 2019.

[37] *Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, Huan Liu.* Fake news detection on social media: A data mining perspective, *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

[38] *Kilian Sprenkamp, Daniel Gordon Jones, Liudmila Zavolokina.* Large language models for propaganda detection, *arXiv preprint* arXiv:2310.06422, 2023.

[39] *Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin.* Attention is all you need, *Advances in Neural Information Processing Systems*, 2017.

[40] *Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela.* Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection, *Annual Meeting of the Association for Computational Linguistics*, 1667–1682, 2021.

[41] *William Yang Wang.* "Liar, liar pants on fire": A new benchmark dataset for fake news detection, *Annual Meeting of the Association for Computational Linguistics*, 422–426, 2017.

[42] *Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou.* MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers, *International Conference on Neural Information Processing Systems*, 5776–5788, 2020.

[43] *Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, Jeffrey Dean.* Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation, *Transactions of the Association for Computational Linguistics*, 5:339–352, 2016.