Multilingual Propaganda Detection: Exploring Transformer-Based Models mBERT, XLM-RoBERTa, and mT5

Mohamed Ibrahim Ragab[1], Ensaf Hussein Mohamed[1], Walaa Medhat[2]
School of Information Technology and Computer Science, CIS,
Nile University, Giza, Egypt
Correspondence: moragab@nu.edu.eg, enmohamed@nu.edu.eg, wmedhat@nu.edu.eg

Abstract

This research focuses on the detection of multilingual propaganda using transformer-based embeddings from state-of-the-art models, including mBERT, XLM-RoBERTa, and mT5. A balanced dataset was employed to ensure equal representation across propaganda classes, enabling robust model evaluation. The mT5 model demonstrated the highest performance, achieving an accuracy of 99.61% and an F1-score of 0.9961, showcasing its effectiveness in multilingual contexts. Similarly, mBERT and XLM-RoBERTa achieved strong results, with accuracies of 92% and 91.41%, respectively, highlighting their capabilities in capturing linguistic and contextual nuances. Despite these high overall performances, the results revealed challenges in detecting subtle propaganda elements, suggesting the need for further improvements in handling nuanced classification tasks.

1 Introduction

Propaganda detection is a critical step in combating the spread of misinformation and biased content designed to manipulate public opinion. Propaganda content often relies on subtle linguistic cues, making its detection a complex task even in monolingual contexts. While recent advancements in natural language processing (NLP), particularly transformer-based models, have significantly improved propaganda detection in English, multilingual detection remains a challenging frontier. This challenge stems from linguistic diversity, contextual variability, and the scarcity of high-quality annotated datasets for non-English languages, which limits the generalizability of existing models.

Previous studies have focused on monolingual approaches, leveraging machine learning and deep learning models to identify and categorize propagandistic content. Transformer-based architectures, such as BERT and its multilingual variants, have demonstrated strong performance in capturing complex linguistic patterns. However, these approaches often face limitations in multilingual settings due to imbalanced datasets, inconsistent performance across languages, and challenges in distinguishing nuanced propaganda categories. Existing models also struggle to balance precision and recall across diverse classes, particularly in low-resource languages.

In this paper, we address these limitations by integrating transformer-based models—mBERT, XLM-RoBERTa, and mT5—into ensemble frameworks designed to enhance multilingual propaganda detection. We aim to improve classification robustness and performance across languages. Our approach leverages the strengths of individual models while mitigating their weaknesses through ensemble learning, offering a more balanced and effective solution for multilingual classification tasks.

The rest of the paper is organized as follows: Section 2 shows related work on the Propaganda classification problem. Section 3.1 describes the dataset and preprocessing techniques used in this study, Section 3.2 details the transformer models and ensemble frameworks implemented. Section 4 presents the experimental results, including individual model performance and ensemble outcomes. Finally, Section 5 discusses the findings, limitations, and potential directions for future research.

2 Related Work

In this section, we review recent advancements in related fields, focusing on propaganda detection, political bias detection, extremism detection, and multilingual misinformation detection. These studies provide a foundation for understanding the challenges and methodologies in multilingual classification tasks, highlighting limitations in existing approaches and motivating the contributions of this work.

The SAFARI (Azizov et al., 2024) study evaluates cross-lingual political bias and factuality

Proceedings of the Proceedings of the 1st International Workshop on Nakba Narratives as Language Resources, pages 75–82
January 20, 2025.

detection using media- and article-level datasets. Ensemble models with soft voting and Multilingual Pre-trained Language Models (MPLMs) like mBERT and XLM-R performed best on English datasets, achieving F1 scores of 84.96% for factuality and 84.95% for political bias. Multilingual datasets showed weaker results due to limited training data, with the best F1 score for political bias reaching 29.05%. Article-level models performed strongly on English-distant supervision datasets (F1: 82.62%) but less so on expert-annotated and multilingual data, highlighting challenges in cross-lingual transfer. Joint modeling, combining political bias and factuality detection, achieved its highest F1 score of 83.81% using soft voting. Large Language Models (LLMs), including Mistral7B and LLaMA27B, underperformed MPLMs in zero-shot settings, with F1 scores up to 46.84%. The study employed clustering techniques for data curation and evaluated models using F1 scores, accuracy, and recall. While ensemble methods and MPLMs proved effective, challenges remain in multilingual adaptation and fine-grained zero-shot learning.

Recent advancements (Modzelewski et al., 2024) in propaganda detection have utilized diverse methods, including fine-tuned transformers, few-shot GPT prompting, and classical machine learning. Studies focused on a dataset of tweets by diplomats from China, Russia, the U.S., and the EU in English and Spanish, tackling binary to fine-grained multilabel classification tasks. Among these, XLM-RoBERTa (XLM-BI) emerged as the top-performing model, excelling in multilingual and Spanish tasks, while RoBERTa (ROBEN) demonstrated strong performance on English-specific tasks. Few-shot prompting with GPT-3.5 and GPT-4o showed potential for binary classification, with GPT-4o outperforming GPT-3.5 but not reaching the accuracy of fine-tuned BERT models. Classical machine learning approaches, including LightGBM and XGBoost with StyloMetrix linguistic features, offered competitive performance, particularly in English binary classification, where LightGBM's F1 score rivaled that of BERT-based models. These findings highlight the strength of fine-tuned transformers for complex multilingual tasks, while also recognizing the effectiveness of classical machine learning and GPT-based methods in specific contexts.

Advancements in extremism and radicalization detection (Zerrouki and Benblidia, 2024) have explored multilingual datasets and sophisticated classification techniques. Recent research introduced a multilingual corpus for binary and multiclass classification tasks, encompassing texts on extremism and radicalization from diverse sources such as ISIS-related content and hate speech in languages like English, Arabic, Indian, Korean, and Kazakh. The dataset includes 17,100 samples for binary extremism, 5,000 for binary radicalization, and 11,400 for multiclass extremism. The study utilized preprocessing methods such as language-specific cleaning, TF-IDF, and word embeddings, alongside machine learning models (e.g., L.SVC, Random Forest) and deep learning approaches (Bi-LSTM, DistilBERT-Multi). Bi-LSTM achieved high accuracy for binary classification (97.8% for radicalization, 96.81% for extremism), while transformer-based models excelled in multiclass classification with 91.07% accuracy. These results highlight the effectiveness of deep learning and transformer models for multilingual extremism detection tasks.

Recent efforts in political bias detection (Agrawal et al., 2022) have introduced an annotated dataset of 1,388 Hindi news articles and headlines from diverse sources, balanced across three categories: biased towards BJP, biased against BJP, and neutral. Articles were annotated for coverage and tonality bias with a kappa score of 0.65, highlighting the subjective challenges in labeling neutrality. The dataset features class-specific averages in word and sentence counts, providing insights into linguistic characteristics. The study evaluated transformer-based models, including mBERT, XLM-RoBERTa, and IndicBERT, alongside traditional machine learning approaches like SVM, Logistic Regression, and Random Forest. XLM-RoBERTa achieved the best results with 83% accuracy and an F1-macro score of 76.4%, significantly outperforming traditional models, which scored below 60% in F1-macro. The findings emphasize the effectiveness of multilingual transformers in bias detection tasks and highlight challenges in accurately identifying neutral articles due to their subjective nature.

Recent research (Panda and Levitan, 2021) on misinformation detection has focused on multilingual datasets and transformer-based models. A study using tweets related to COVID-19 from the NLP4IF 2021 shared task investigated misinformation detection in English, Bulgarian, and Arabic. The dataset included binary annotations for seven questions assessing misinformation charac-

teristics, with 451–3,000 training samples and up to 1,000 test samples per language. The study evaluated logistic regression, a transformer encoder, and BERT-based models. English BERT (Capuozzo et al., 2020) achieved the best results for English (F1: 0.729), while multilingual BERT (m-BERT) demonstrated strong cross-lingual generalization, achieving F1 scores of 0.843 for Bulgarian and 0.741 for Arabic. Experimental setups, including zero-shot, few-shot, and target-only training, highlighted the potential of m-BERT to perform well in low-resource settings with minimal target-language data. These findings emphasize the effectiveness of contextualized embeddings and multilingual transformers in detecting misinformation across diverse languages.

Gap Analysis of Related Work: Despite the advancements in multilingual propaganda detection, several gaps remain that highlight the need for further exploration:

Limited Multilingual Representation: While studies like SAFARI and NLP4IF have explored multilingual contexts, their datasets are often limited in linguistic diversity, focusing on high-resource languages like English and Spanish. Low-resource languages, which are equally vulnerable to propaganda and misinformation, remain underrepresented, impacting the generalizability of existing models.

Overreliance on Individual Models: Transformer models like mBERT, XLM-RoBERTa, and mT5 have shown strong standalone performance, but their reliance on pretraining data biases can limit robustness in real-world scenarios. Ensemble methods are underexplored in mitigating these weaknesses, particularly in balancing class-specific performance.

3 Materials and Methods

This section describes the methodology for detecting multilingual propaganda using transformer-based embeddings and ensemble learning. The process involves balancing the dataset to ensure equal class representation and preprocessing steps like tokenization. Transformer models—mBERT, XLM-RoBERTa, and mT5—were fine-tuned to extract multilingual embeddings, capturing linguistic and contextual nuances. These embeddings were then used in classification frameworks to evaluate the models' effectiveness in handling multilingual propaganda detection tasks.

3.1 The used Dataset

The dataset (Duaibes et al., 2024) utilized in this study, curated by SinaLab (2024), consists of a comprehensive collection of 13,500 rows and 13 columns, representing a rich and diverse array of Facebook posts. Developed as part of the FigNews 2024 Shared Task on News Media Narratives, the dataset focuses on the framing of the Israeli War on Gaza, providing valuable insights into bias and propaganda in media. The corpus spans five languages—Arabic, English, Hebrew, French, and Hindi—with an equal distribution of 2,400 posts per language, ensuring linguistic diversity and balance. Each post is meticulously annotated for attributes related to bias and propaganda, making this dataset a critical resource for advancing multilingual analysis of media narratives.

The dataset includes the following notable columns: Text is the original text of the Facebook post, as it appeared in its source language. English MT is a machine-translated version of the text in English, facilitating cross-lingual analysis and ensuring uniformity for annotators who are not fluent in the source language. Arabic MT is a machine-translated version of the text in Arabic, aiding linguistic diversity and analysis. Propaganda This column indicates whether the post contains propaganda types.

A particular focus of our analysis is the Propaganda column, which captures whether a post contains propaganda. As shown in Table 1 and Figure 1, this column consists of four distinct classes, with the distribution of instances as follows:

Class	Number of Instances
Not Propaganda	7098
Propaganda	3301
Unclear	269
Not Applicable	132

Table 1: Distribution of instances in the Propaganda column across the dataset.

These labels provide a foundation for investigating the prevalence and characteristics of propaganda within the dataset. The predominance of "Not Propaganda" suggests that most posts lack propagandistic content, yet the substantial presence of "Propaganda" emphasizes the significance of its impact in framing narratives. The smaller proportions of "Unclear" and "Not Applicable" highlight the challenges and ambiguities faced during anno-

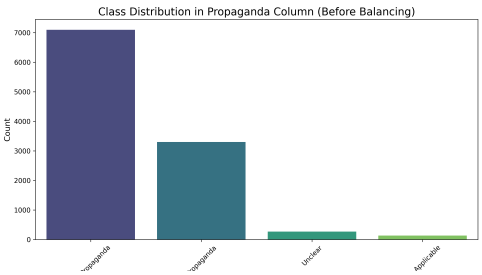

Figure 1: Distribution of instances in the Propaganda column across the dataset.

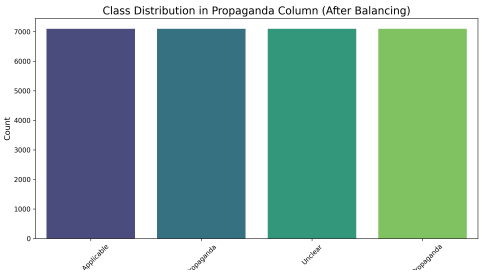

Figure 2: Class Distribution in the Propaganda Column After Applying the Balancing Step.

tation.

Class	Number of Instances
Not Propaganda	7098
Propaganda	7098
Unclear	7098
Not Applicable	7098

Table 2: Class Distribution in the Propaganda Column After Applying the Balancing Step.

As shown in Table 2 and Figure 2, After applying the balancing step to the dataset, each class in the Propaganda column was adjusted to have an equal number of instances, with 7,098 samples per class. This ensured that the dataset was balanced, eliminating any bias towards overrepresented classes and providing a more even distribution for training the model.

The dataset also includes metadata such as the language of the post, bias-related annotations, and machine-translated versions of the text for cross-lingual analysis. These features enable a comprehensive exploration of linguistic and cultural patterns in bias and propaganda.

Figure 3 provides a snapshot of the dataset used in this study. It showcases the key columns, including the Text column containing the multilingual propaganda content and the Propaganda column indicating class labels. The figure highlights the structure of the data, demonstrating how text samples are paired with corresponding annotations, which serve as the basis for training and evaluating the models.

	Batch	Source Language	ID	Type	Text	English MT	Arabic MT	Annotator ID	Bias	Propaganda	Type of Propaganda	Type of Bias	Comments
0	B01	English	1	MAIN	Yemen's Houthis have waded into the Israel-Ham...	Yemen's Houthis have waded into the Israel-Ham...	خاض الحوثيون في اليمن الحرب بين إسرائيل وحماس ...	1.0	Biased against Palestine	Not Propaganda	Propaganda Not to be deleted	ضمني	NaN
1	B01	English	2	MAIN	Isreal - Hamas Conflict | Face to Face	Isreal - Hamas Conflict | Face to Face	إسرائيل - الصراع مع حماس | وجها لوجه	4.0	Unbiased	Not Propaganda	Not Propaganda	NaN	NaN

Figure 3: An Overview of the Dataset: Sample Columns and Rows

3.2 Methodology

In this study, we implemented state-of-the-art multilingual models to classify propaganda and bias in Facebook posts. The models include mBERT, mT5, and XLM-RoBERTa.

As shown in Figure 4, the proposed architecture for multilingual propaganda detection leverages XLM-Roberta, a powerful multilingual transformer model, combined with robust preprocessing and training strategies. The first step involves preparing the dataset by loading text data in multiple languages (e.g., English, and Arabic) and their corresponding propaganda labels. To address the class imbalance, oversampling is applied to minority classes to ensure balanced representation across all categories. Text data is preprocessed by removing punctuation and converting to lowercase to standardize inputs, followed by label encoding to convert categorical labels into numerical values.

The next stage utilizes the Model tokenizer to tokenize the preprocessed text while applying padding and truncation to ensure uniform input lengths. A HuggingFace dataset is created and split into training and testing subsets. The architecture uses the MBert, XLM-Roberta, and MT5) models for sequence classification, configured for four output classes corresponding to the encoded labels. Training configurations include a low learning rate, small batch size, weight decay for regularization, and an evaluation strategy that monitors performance after each epoch.

Training is conducted with a HuggingFace Trainer, integrating a manual early stopping mechanism to prevent overfitting. The model evaluates validation loss at the end of each epoch, saving

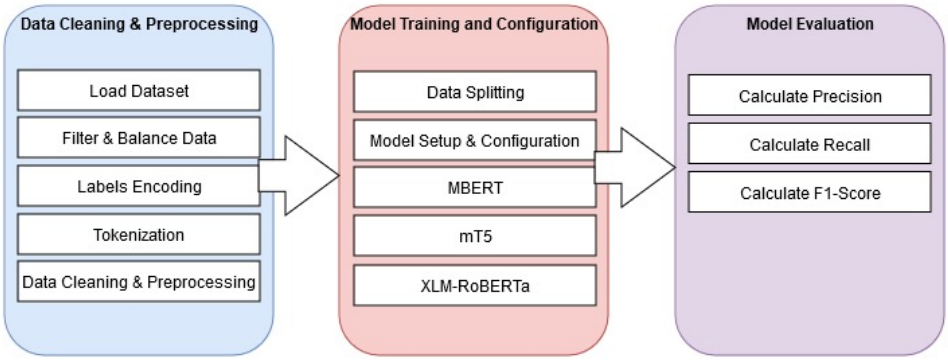

Figure 4: Multilingual Propaganda Detection Architecture Utilizing XLM-Roberta, MBert, and MT5 with Balanced Class Handling and Early Stopping for Optimized Performance.

the best-performing model when improvements are detected and halting training if no improvement occurs for a specified number of consecutive epochs. Finally, the best model is evaluated on the test set, with metrics such as accuracy, precision, recall, and F1-score reported to assess classification performance. This architecture effectively combines XLM-Roberta's multilingual capabilities with balanced data handling and efficient training strategies to achieve robust propaganda detection.

Multilingual BERT (MBERT) is a variant of BERT pre-trained on Wikipedia data across 104 languages, as described by Libovický et al. (2019). Designed to process input at the token level, it utilizes deep bidirectional attention mechanisms to capture intricate linguistic relationships and contextual nuances effectively. Its multilingual pretraining makes it particularly suited for tasks involving cross-lingual understanding and classification, rendering it an ideal choice for analyzing diverse datasets like the one used in this study. mBERT's versatility has been demonstrated across a wide range of multilingual and cross-lingual natural language processing applications.

mT5 (Multilingual T5) is an extension of the T5 model, pre-trained on over 100 languages, as described by Fuadi et al. (2023). It employs a text-to-text framework, where all tasks—ranging from classification to text generation—are reformulated as text generation problems. This unified approach allows mT5 to handle a wide variety of language processing tasks with remarkable flexibility and consistency. Its multilingual training on diverse linguistic data makes it particularly robust, even in low-resource language settings, enabling effective performance across a broad spectrum of multilingual and cross-lingual tasks.

XLM-RoBERTa It builds on RoBERTa, an optimized version of BERT (Wiciaputra et al., 2021), by pretraining on a massive multilingual corpus spanning 100 languages. It achieves state-of-the-art performance on various multilingual benchmarks and is particularly effective in handling languages with limited resources.

4 Results and Discussion

4.1 Experiment Setup

We employed three transformer-based models — **mBERT**, **mT5**, and **XLM-RoBERTa** — for multilingual propaganda detection. The experiment setup consisted of data preprocessing, model fine-tuning, and evaluation.

4.1.1 Computational Resources

The experiments were conducted on the Kaggle platform, leveraging a P100 GPU for accelerated computations. The P100 GPU provided the necessary computational power for handling large datasets and fine-tuning transformer-based models efficiently. The use of a high-performance GPU significantly reduced training time, especially for resource-intensive models like mT5. Kaggle's environment also facilitated seamless access to datasets and libraries required for the experiments.

4.1.2 Data Preprocessing

- **Cleansing**: Rows with missing values were removed to ensure data consistency.

- **Class Balancing**: Oversampling techniques were employed to balance the number of instances across all classes, mitigating potential bias during training.

- **Tokenization**: Each model utilized its respective tokenizer - mBERT: `BertTokenizer` - mT5: `T5Tokenizer` - XLM-RoBERTa: `XLM-RoBERTaTokenizer` The text data was tokenized into subword units, ensuring compatibility with the transformer architectures.
- **Label Encoding**: Categorical labels, such as "Propaganda" and "Not Propaganda," were numerically encoded to facilitate classification.
- **Dataset Splits**: The dataset was converted into PyTorch tensors and split into training, validation, and test sets. An 80-10-10 split ratio was maintained to balance the training and evaluation phases.

4.1.3 Model Fine-Tuning

- **mBERT**: `BertForSequenceClassification`, learning rate 2×10^{-5}, batch size: 16, epochs: 50 (early stopping applied), no. of labels: 4.
- **mT5**: `MT5ForConditionalGeneration`, learning rate 5×10^{-5}, batch size: 8, epochs: 50 (early stopping applied), no. of labels: 4.
- **XLM-RoBERTa**: HuggingFace `Trainer`, learning rate 1×10^{-5}, batch size: 8, epochs: 50 (early stopping applied), no. of labels: 4.

All models were optimized using the AdamW optimizer, with weight decay set to 0.05. Early stopping was implemented for all models to prevent overfitting, with training halting after three consecutive epochs without validation loss improvement. Learning rate schedulers were utilized to adjust learning rates dynamically during training.

4.2 Result

In the Results section, we analyze and compare the performance of three transformer-based models—mBERT, MT5, and XLM-RoBERTa—on the task of multilingual propaganda detection. The analysis includes a detailed evaluation of the models' performance with and without data balancing techniques to address class imbalances. Key evaluation metrics, including accuracy, precision, recall, F1-score, training loss, validation loss, and the number of training epochs, are used to assess and compare the effectiveness of each model under both scenarios.

Performance on Imbalanced Data (Without Balancing) As shown in Table 3, the performance of the models varied significantly when trained on imbalanced data, highlighting the challenges posed by class distribution. The MT5 model achieved a remarkable accuracy of 98.86%, precision of 0.9911, recall of 0.9886, and F1-score of 0.9882, showcasing its ability to handle imbalanced datasets effectively. This superiority can be attributed to MT5's text-to-text framework, which allows it to model nuanced relationships within the data.

In contrast, mBERT struggled with imbalanced data, achieving an accuracy of only 53.00%, with precision, recall, and F1-score all around 0.65. This indicates difficulty in generalizing across uneven class distributions. Similarly, XLM-RoBERTa achieved moderate performance, with an accuracy of 69.70% and an F1-score of 0.5595, suggesting it was able to identify positive instances (recall of 0.6870) but lacked precision (0.4720). These results underscore the necessity of addressing class imbalances to improve model performance.

Performance on Balanced Data (After Balancing) After addressing class imbalances through oversampling techniques, all models showed significant improvements in their performance metrics, as detailed in Table 4. The MT5 model continued to outperform the other models, achieving an accuracy of 99.61%, precision of 0.9961, recall of 0.9961, and F1-score of 0.9962. This further highlights MT5's robustness in handling balanced datasets and extracting nuanced features from multilingual text.

The mBERT model demonstrated substantial improvement, achieving an accuracy of 92.0%, with balanced precision, recall, and F1-scores of 0.92. Its ability to generalize effectively after balancing emphasizes the importance of preprocessing in enhancing performance. XLM-RoBERTa, while still lagging behind MT5 and mBERT, improved to an accuracy of 89.51%, with an F1-score of 0.8934, indicating better adaptation to the balanced dataset.

Training and Validation Loss Table 5 provides insights into the training and validation loss for each model. The MT5 model exhibited the most efficient learning, achieving the lowest training loss of 0.0080 and validation loss of 0.0102, highlighting its excellent generalization capabilities. The mBERT model followed with a training loss of 0.0755 and validation loss of 0.2719, reflecting steady learning and robust generalization. XLM-RoBERTa, however, recorded a training loss of 0.0467 but struggled with a validation loss of 0.7156, indicating potential overfitting or difficulty in adapting to the multilingual dataset's complex-

Model	Accuracy (%)	Precision	Recall	F1-Score	No. Epocs
mBERT	53.00	0.66	0.64	0.65	5
MT5	98.86	0.9911	0.9886	0.9882	24
XLM-RoBERTa	69.70	0.4720	0.6870	0.5595	15

Table 3: Performance Metrics of mBERT, MT5, and XLM-RoBERTa Models Without Balancing Techniques.

Model	Accuracy (%)	Precision	Recall	F1-Score	No. Epocs
mBERT	92.0	0.98	0.92	0.92	8
MT5	99.61	0.9961	0.996	0.9962	27
XLM-RoBERTa	89.51	0.9006	0.8951	0.8934	16

Table 4: Performance Metrics of mBERT, MT5, and XLM-RoBERTa Models Using Balancing Techniques.

ity.

Model	Training	Validation
mBERT	0.0755	0.2719
MT5	0.0080	0.0102
XLM-RoBERTa	0.0467	0.7156

Table 5: Training and Validation Loss of mBERT, MT5, and XLM-Roberta Models.

Analysis of Performance Differences: MT5's Dominance: MT5 consistently outperformed the other models across all metrics, both with and without data balancing. Its text-to-text framework enables it to capture subtle linguistic and contextual nuances, making it highly effective for multilingual propaganda detection, making it highly effective for multilingual propaganda detection with an accuracy of 99%.

mBERT's Consistency: Despite not achieving the same level of accuracy as MT5, mBERT demonstrated commendable performance with an accuracy of 92% and competitive scores across all metrics. Its ability to leverage multilingual pretraining makes it a robust choice for tasks where efficiency is prioritized over peak accuracy.

Challenges with XLM-Roberta: While XLM-RoBERTa showed promise with an accuracy of 89.51%, its higher validation loss suggests issues with overfitting or insufficient adaptation to the dataset's multilingual nature. This could stem from its reliance on pretraining that may not fully capture the nuanced biases present in propaganda detection tasks.

Generalization and Efficiency: The results highlight the importance of achieving a balance between learning efficiency and generalization. While all models employed early stopping to mitigate overfitting, MT5's consistently low validation loss underscores its superior ability to generalize, whereas XLM-RoBERTa's performance suggests room for improvement in adapting to diverse linguistic inputs.

The analysis reveals that data balancing significantly enhances model performance, especially for models like mBERT and XLM-RoBERTa, which struggled with imbalanced datasets. MT5's consistently high performance underscores its suitability for multilingual tasks, while the improvements seen in mBERT demonstrate its potential when coupled with effective preprocessing techniques. These findings emphasize the critical role of balancing and preprocessing in ensuring fair and robust evaluations.

Overall, these findings demonstrate that while MT5 is the most effective model for this task, further research could focus on improving the generalization capabilities of models like XLM-RoBERTa to better handle the complexities of multilingual propaganda detection.

5 Conclusion

This study implemented and evaluated three transformer-based models—mBERT, XLM-RoBERTa, and mT5—to address the challenge of multilingual propaganda detection. By leveraging these models' multilingual capabilities and employing advanced preprocessing techniques, including data balancing, we conducted a comprehensive assessment of their performance. Key evaluation metrics such as accuracy, precision, recall, F1-score, and loss values were used to determine the models' effectiveness across diverse propaganda categories.

The mT5 model consistently outperformed its counterparts, achieving an outstanding accuracy

of 99.61% and an F1-score of 0.9961, demonstrating its exceptional ability to handle multilingual content and detect propaganda with high precision. Its text-to-text framework allowed it to effectively model linguistic nuances across multiple languages, making it the most reliable model for this task. The mBERT model also showcased strong performance, achieving an accuracy of 92.0% and an F1-score of 0.92, excelling in the "Not Propaganda" and "Not Applicable" categories. Despite these results, it showed room for improvement in more nuanced categories. The XLM-RoBERTa model achieved a respectable accuracy of 89.51%, with an F1-score of 0.8934, but faced challenges with generalization, as evidenced by its higher validation loss compared to the other models.

The results emphasize the transformative potential of transformer-based embeddings in multilingual propaganda detection. While mT5 emerged as the most effective model, mBERT demonstrated computational efficiency and solid performance, making it a viable choice for practical applications. On the other hand, XLM-RoBERTa highlighted areas for future improvement, particularly in adapting to complex multilingual tasks.

Limitations and Future Work: These future directions aim to refine and enhance the capabilities of multilingual propaganda detection, expanding the models' adaptability, accuracy, and generalization across various contexts and languages.

Hyperparameter Optimization and Fine-Tuning: Future work could explore the fine-tuning of hyperparameters such as learning rate, batch size, and number of epochs for each model, especially mBERT and XLM-RoBERTa. Optimizing these parameters could lead to improved performance, particularly in terms of accuracy, precision, and recall across different propaganda categories.

Exploring Advanced Ensemble Techniques: While individual models like mT5, mBERT, and XLM-RoBERTa demonstrated strong performance, future research could investigate the use of more advanced ensemble methods, such as stacking and boosting, to combine the strengths of multiple models. This could help in improving performance, especially in identifying subtle propaganda elements that individual models may miss.

Cross-Lingual Transfer Learning: One promising direction is to explore cross-lingual transfer learning by leveraging pre-trained multilingual models to better handle low-resource languages.

6 References

Samyak Agrawal, Kshitij Gupta, Devansh Gautam, and Radhika Mamidi. 2022. Towards detecting political bias in hindi news articles. In *Proceedings of the 60th annual meeting of the association for computational linguistics: student research workshop*, pages 239–244.

Dilshod Azizov, Zain Mujahid, Hilal AlQuabeh, Preslav Nakov, and Shangsong Liang. 2024. Safari: Cross-lingual bias and factuality detection in news media and news articles. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12217–12231.

Pasquale Capuozzo, Ivano Lauriola, Carlo Strapparava, Fabio Aiolli, Giuseppe Sartori, et al. 2020. Automatic detection of cross-language verbal deception. In *42nd Annual Conference of the Cognitive Science Society (CogSci'20)*, pages 1756–1762.

Lina Duaibes, Areej Jaber, Mustafa Jarrar, Ahmad Qadi, and Mais Qandeel. 2024. Sina at FigNews 2024: Multilingual Datasets Annotated with Bias and Propaganda. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*, Bangkok, Thailand. Association for Computational Linguistics.

Mukhlish Fuadi, Adhi Dharma Wibawa, and Surya Sumpeno. 2023. Adaptation of multilingual t5 transformer for indonesian language. In *2023 IEEE 9th Information Technology International Seminar (ITIS)*, pages 1–6. IEEE.

Jindřich Libovickỳ, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual bert? *arXiv preprint arXiv:1911.03310*.

Arkadiusz Modzelewski, Paweł Golik, and Adam Wierzbicki. 2024. Bilingual propaganda detection in diplomats' tweets using language models and linguistic features. *IberLEF@ SEPLN*.

Subhadarshi Panda and Sarah Ita Levitan. 2021. Detecting multilingual covid-19 misinformation on social media via contextualized embeddings. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 125–129.

Yakobus Keenan Wiciaputra, Julio Christian Young, and Andre Rusli. 2021. Bilingual text classification in english and indonesian via transfer learning using xlm-roberta. *International Journal of Advances in Soft Computing & Its Applications*, 13(3).

Khadidja Zerrouki and Nadjia Benblidia. 2024. Multilingual text preprocessing and classification for the detection of extremism and radicalization in social networks.