PolyTruth: Multilingual Disinformation Detection using Transformer-Based Language Models

Anonymous*, Anonymous*, Anonymous*

Anonymous*, Anonymous*, Anonymous*

Abstract. Disinformation spreads rapidly across linguistic boundaries, yet most AI models are still benchmarked only on English. We address this gap with a systematic comparison of five multilingual transformer models: mBERT, XLM, XLM-RoBERTa, RemBERT, and mT5 on a common fake-vs-true machine learning classification task. While transformer-based language models have demonstrated notable success in detecting disinformation in English, their effectiveness in multilingual contexts still remains up for debate. To facilitate evaluation, we introduce PolyTruth Disinfo Corpus, a novel corpus of 60,486 statement pairs (false claim vs. factual correction) spanning over twenty five languages that collectively cover five language families and a broad topical range from politics, health, climate, finance, and conspiracy, half of which are fact-checked disinformation claims verified by an augmented MindBugs Discovery dataset. Our experiments revealed significant performance variations. Models such as RemBERT achieved superior overall accuracy, particularly excelling in low-resource languages, whereas models like mBERT and XLM exhibit considerable limitations when training data is scarce. We provide a discussion of these performance patterns and implications for real-world deployment. The data set will also be publicly available on our GitHub repository to encourage further experimentation and advancement in multilingual disinformation detection. Our findings illuminate both the potential and the current limitations of AI systems for disinformation detection.

Keywords: Disinformation Detection · Multilingual NLP · Transformer Models · PolyTruth Disinformation Corpus · Cross-Lingual Evaluation

1 Introduction

Online disinformation has become pervasive, spreading rapidly across social media and online platforms. Studies have shown that false news can spread faster and farther than the truth in online networks [4], exacerbating the challenge for automated systems. Disinformation detection has thus attracted significant research attention in recent years [1]. However, much of the earlier work focused on English-language content, taking advantage of large labeled datasets and advances in deep learning for text classification [2]. The multilingual dimension of

the disinformation problem remains relatively under-explored due to the scarcity of annotated data in languages beyond English [11, 12]. One of the broad motivations of this study was the Facebook (Meta) scandal in Myanmar and system-level inability of its algorithms to filter out harmful content because it didn't have enough training data on the language [5, **?**]. There are many reasons for this, but one reasons is that building robust multilingual disinformation detectors is inherently difficult, and false claims often propagate in multiple languages, including low-resource languages where detection tools are limited [13]. Multilingual transformer-based language models offer a potential solution for this task. Models such as mBERT [6] and XLM-RoBERTa [8] learn joint representations for dozens of languages and have demonstrated impressive cross-lingual transfer capabilities on tasks like question answering. Recent studies have started to explore multilingual detection using such models [12], and have found that multilingual training strategies can substantially improve performance. However, in the literature, comparisons of different multilingual transformers on a unified disinformation detection task are lacking. In this paper, we address this gap by comparing five state-of-the-art multilingual transformer models on the task of disinformation statement classification. We make use of the MindBugs Discovery dataset [14], which contains 30,243 debunked disinformation statements from Europe (2009–2024) in various languages. Each statement is annotated with its veracity (false/disinformation) and language. We augmented and created a complimentary dataset with true (non-disinformation) statements to enable binary classification.[1] We fine-tune and evaluate five transformer-based language models: BERT-base-multilingual-cased (mBERT) [6], XLM (Cross-lingual Language Model) [7], XLM-RoBERTa (base) [8], RemBERT [9], and mT5 [10] under the same experimental conditions. Our evaluation focuses on two key aspects, the overall model performance on the multilingual data, and performance differences between high-resource and low-resource language subsets.

Our contributions are as follows: **(1)** We expand prior work by conducting a thorough comparative evaluation of five multilingual transformer models on a common disinformation detection task, providing insights into their strengths and weaknesses across languages. **(2)** We describe a unified training and evaluation framework for multilingual fake news detection, including dataset preparation and preprocessing steps that can serve as a reference for future research. **(3)** We analyse model performance across languages, highlighting how high-resource languages benefit from abundant data and how low-resource languages pose challenges where certain models (notably XLM-R and RemBERT) still perform robustly. To our knowledge, this is a comprehensive evaluation of multilingual transformers for disinformation detection. We hope our findings will inform the development of more effective multilingual and cross-lingual fake news detection systems.

[1] For each false claim, we paired it with a corresponding true statement or a factual correction from fact-checking sources to serve as a negative example. This yielded a balanced dataset for training.

The remainder of the paper is organised as follows. Section 2 reviews related work in multilingual disinformation detection. Section 3 describes the dataset and our methodology, including preprocessing and model fine-tuning setup. Section 4 presents the experimental results, comparing model performances overall and by language groups, with analysis. Finally, Section 5 concludes the paper and outlines future work.

2 Related Work

Early research on disinformation and fake news detection largely centered on English content, using machine learning and deep learning to classify news articles or social media posts as fake or real [1]. Traditional approaches relied on textual features like n-grams, readability, and linguistic cues or even incorporated user and network metadata. With the rise of deep neural models, many works applied CNNs, RNNs, or attention-based architectures to detect disinformation, showing improved accuracy on benchmark datasets. However, these efforts often did not generalise to other languages due to language-specific patterns and resource disparities. Addressing multilingual detection presents a unique challenge because labeled data in most languages are scarce. One line of work has focused on cross-lingual transfer, where a model trained on a high-resource language is applied to detect fake news in low-resource languages. Studies in the past have translated non-English data into English to use existing detectors, or conversely translated English training data into the target language to fine-tune models [12]. While these translation-based strategies leverage English resources, they risked meaning loss and translation errors, especially for low-resource languages. Another approach was to use multilingual transformers pre-trained on multiple languages. These models could directly encode text in different languages into a shared semantic space, enabling zero-shot or few-shot transfer. The multilingual BERT model (mBERT) [6] was one of the first to show cross-lingual capabilities, despite being trained without explicit alignment objectives, mBERT can often generalise to languages unseen in fine-tuning. Lample and Conneau's XLM model [7] improved cross-lingual performance by introducing a translation language modeling objective to better align representations across languages. The XLM-RoBERTa (XLM-R) model [8] further advanced the state of the art by training on a massive multilingual corpus known as Common Crawl data containing 100 languages, and with a larger model size and improved pre-training, this led to significant gains on tasks like cross-lingual natural language inference and question answering.

These multilingual transformers have been applied to disinformation tasks in recent years and challenges have been put out to the NLP communities to address this. The CheckThat! 2022 Lab included a cross-lingual fake news detection task (English to German) where participants successfully fine-tuned XLM-R for transfer learning [15]. This built on Patwa et al. [16], in the CONSTRAINT 2021 shared task on fake news detection in English and Hindi, demonstrating that transformers can handle multilingual and even code-switched content

given appropriate fine-tuning. Further research conducted on various models and training scenarios [12] experimented with monolingual vs. multilingual training, noting that training on combined multilingual data can improve low-resource language performance, albeit with some trade-offs for high-resource languages due to class imbalance. In follow-up work, Chalehchaleh et al. [17] also looked at large language model-based data augmentation to enhance multilingual fake news detection by generating synthetic training examples for low-resource languages using GPT-style models, reporting improved performance and highlighting data augmentation as a viable solution to the low-resource problem. This is in line with our work. Lesser explored advancements have also looked the integration of external evidence into multilingual disinformation detection systems. Hammouchi and Ghogho [13] proposed an evidence-aware multilingual fake news detection framework, which combines multilingual transformer models with evidence retrieval and source credibility mechanisms. This approach, evaluated on COVID-19 disinformation across multiple languages, showed significant accuracy improvements (e.g., achieving an F1-score of 0.85 on the XFACT dataset). Another method shown by Zhou et al.[18] explored cross-lingual knowledge transfer techniques, demonstrating that leveraging related languages during fine-tuning can substantially enhance performance on low-resource language disinformation tasks. A final promising direction is the use of integrated multi-modal signals [19], such as combining images and text, for multilingual fake news detection, this demonstrated that multimodal models can significantly outperform text-only baselines, particularly useful as they could be employed by very large online platforms due to how visually rich they are.

Despite these developments, the literature still lacks extensive head-to-head comparisons of multiple multilingual transformer models on disinformation detection tasks. Most existing studies evaluate a single model or only a few models in isolation. We aim to address this gap by systematically evaluating five multilingual transformers under identical conditions, providing insights into their relative strengths and limitations. Our work differs from evidence-based approaches in that we focus on content-based detection, where the models make predictions solely from the statement text. This setup allows us to directly assess the language understanding and generalisation capability of multilingual transformers on the task. This can also help identify which architectures are most suitable for multilingual disinformation detection on similar datasets.

3 Data and Methodology

3.1 Dataset and Preparation

We base our experiments on a multilingual disinformation dataset provided as part of the MindBugs Discovery project [14]. The dataset consists of 30,243 statements that have been identified as disinformation (fake or misleading claims) by verified fact-checking organizations in Europe between 2009 and 2024. Each data instance includes the text of the statement, the language of the statement, the

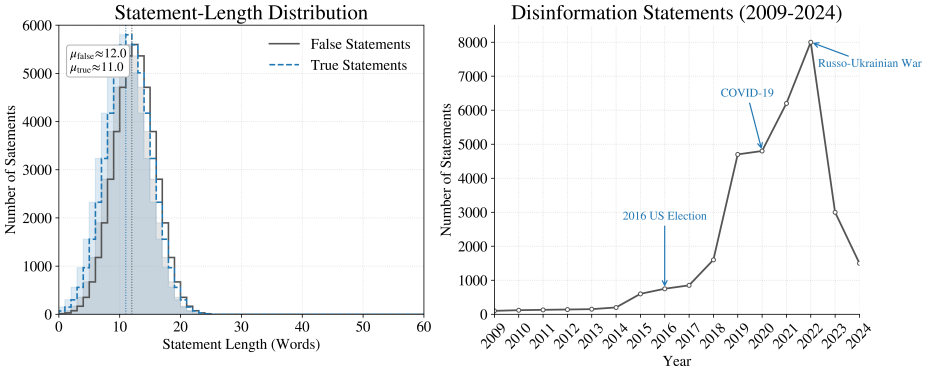

Fig. 1. Left—**Statement length**: both false and true claims peak at 10–15 words; false statements exhibit a slightly longer tail. Right—**Yearly volume**: counts stay low until 2015, rise with the *2016 US election*, spike during the *COVID-19* infodemic (2020) and surge again with the *2022 Russo-Ukrainian war*.

date it was debunked, and other metadata such as the fact-check source. The statements cover a wide range of topics from political falsehoods, health disinformation to conspiracy theories, which is reflective of real-world disinformation content circulating in different countries and languages. Figure 1 shows some insights into this dataset.

For our task, we treat fake (disinformation) detection as a binary classification problem, given a statement, predict whether it is *disinformation* or *legitimate*. Since the collected statements are all debunked false claims, we constructed a complementary set of true or factual statements to act as the negative class called PolyTruth. Specifically, for each false statement in the dataset, we generated a corresponding true statement using OpenAI's API. These state the corrected fact, or a factual claim from a reliable source on the same topic, creating a counter-claim to the disinformation statement. In cases where a direct corrected statement was not available, we selected a truthful statement from a news articles in the same language and topic. This process resulted in a balanced dataset with an equal number of true vs. false statements. After data augmentation and pairing, our full dataset for training and evaluation comprised approximately 60,486 statements (half disinformation, half true information) across all languages, consisting of 25+ distinct languages in the dataset. Table 1 shows an example of some statements.

The dataset includes a diverse representation of languages, with Russian, Portuguese, German, and Czech being the most represented languages, each containing over 3,000 statements. Notably, Russian has the greatest representation, with approximately 4,488 false statements and a corresponding number of true pairs. Other languages such as Dutch, French, Spanish, and English each contribute between 1,500 to 3,000 statements. Several medium-resource languages like Polish, Arabic, Hungarian, and Romanian have around 1,000 statements

Table 1. Sample *PolyTruth* statements in six languages

Language	**Disinformation statement**	**Corrective (true) statement**
English	BBVA texts customers, asking them to click a link to unblock "suspicious activity".	BBVA never requests unblocking via links; such SMS are confirmed phishing attempts.
Spanish	"España prohibirá el uso de hornos de gas en hogares a partir de 2025".	El Gobierno sólo estudia ayudas para electrodomésticos eficientes; no existe tal prohibición.
German	Die WHO plane ab 2024 einen weltweiten Impfpass, sonst seien Reisen verboten.	Es gibt nur eine Empfehlung für freiwillige digitale Zertifikate; ein Reiseverbot ist nicht beschlossen.
Romanian	Guvernul va introduce o taxă de 5 % pentru fiecare tranzacție bancară online în 2025.	Ministerul Finanțelor dezminte: nu există niciun proiect de lege pentru o astfel de taxă.
Estonian	Tallinna kraanivesi sisaldab ohtlikku pliid; enne joomist tuleks vesi keeta.	Tallinna Vesi ja Terviseamet kinnitavad, et pliisisaldus on normi piires ja vesi on joogikõlbulik.
Latvian	Latvija atcels eiro un 2025. gadā ieviesīs jaunu digitālo valūtu "LVD".	Finanšu ministrija skaidro, ka nekādu plānu par eiro atcelšanu un "LVD" ieviešanu nav.

each. Conversely, the dataset also includes several low-resource languages such as Azerbaijani, Estonian, and Latvian, each having fewer than 500 statements. This language distribution is visualised in Figure 3 provided dataset figures. This imbalance provides an opportunity to examine how well models trained on the pooled data can handle languages with limited training samples. Figure 2 provides the full architecture of our methodology. All text was lowercased and stripped of any URLs or user mentions, in the case of social media-originated statements, as a basic preprocessing step. We did not remove stopwords or punctuation, since modern transformer models can handle raw text and often benefit from the presence of natural language cues. Non-English scripts were left as-is, as the multilingual models can encode them. We did, however, replace any occurrences of explicit fact-check verdict phrases like "(False)" or "(Hoax)" tags sometimes appended to statements in sources to avoid giving clues to the model. After cleaning, each statement was tokenised using the subword tokeniser of the respective model during fine-tuning. We split the data into training, validation, and test sets. To ensure evaluation across languages, we adopted a stratified sampling strategy where the data was partitioned such that each language is represented in all three sets. We used an 80/10/10 split. The final training set contained about 48k statements, with the remainder split evenly into validation (6k) and test (6k). Importantly, the class balance (fake vs. true) was maintained in each subset, and no statements from the same fact-check report were allowed to split across train and test to prevent near-duplicates or memorisation.

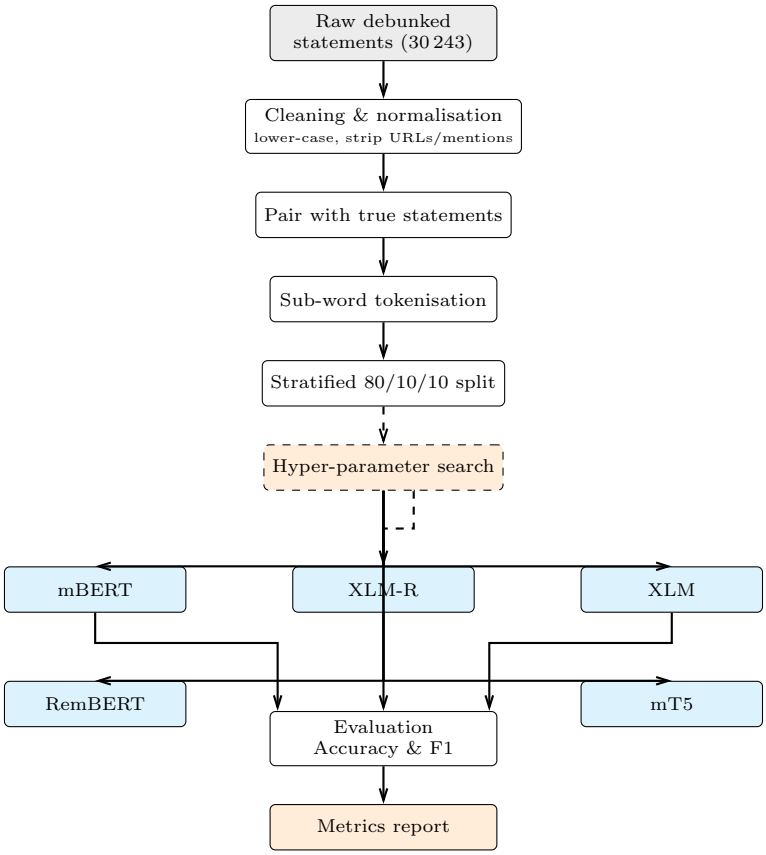

Fig. 2. End-to-end pipeline for multilingual disinformation detection.

3.2 Transformer models evaluated

We fine-tuned five multilingual transformers under identical settings (binary *true* vs. *false* objective). Table 2 shows the architecture of the evaluated models. The encoder-only models (mBERT, XLM-100, XLM-R, RemBERT) use a single linear classifier on the `[CLS]` token, whereas the encoder–decoder model mT5 is trained to emit the word *"true"* or *"false"*.

All models were initialised from publicly available checkpoints in the HuggingFace Transformers library. Input sequences were capped at 512 sub-word tokens; the average statement length in our data is ∼20 tokens, so truncation was negligible. Vocabulary sizes range from 119 M sub-words (mBERT) to 250 k (XLM-R, RemBERT, mT5), ensuring coverage of every language in the corpus. All models were implemented using the HuggingFace Transformers library. We initialised each with the pre-trained weights provided by their authors. Notably, the five models cover a range of model sizes (110M up to 580M), architectures (encoder-only vs. seq2seq), and training regimes (with or without

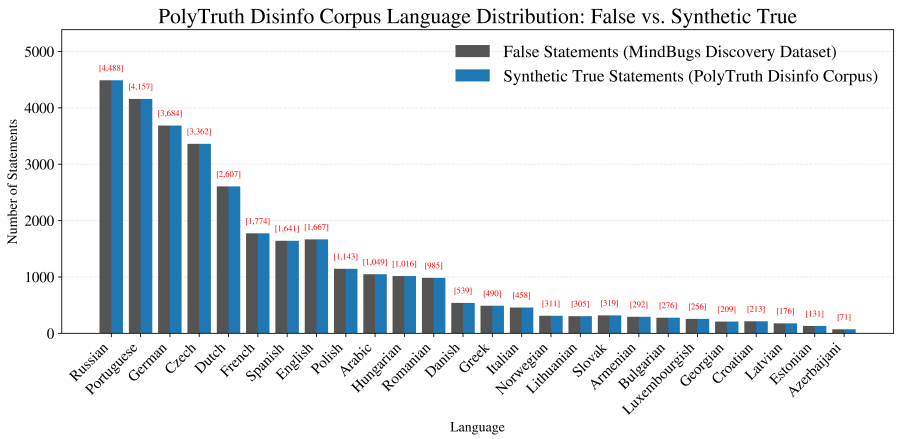

Fig. 3. PolyTruth disinformation corpus: number of *false* (MindBugs) and *synthetic true* statements per language. The five largest languages are Russian, Portuguese, German and Czech each exceed 3,000 statements, while languages such as Latvian, Slovak, Estonian and Azerbaijani have less than a few hundred. Thus highlights the pronounced imbalance across languages. We regard the ten languages with at least 1,000 statements (Russian through Hungarian) as *high-resource*. All remaining languages, from Romanian down to Azerbaijani, form the *low-resource* group. This division underpins our subsequent analysis of how model performance varies with data availability.

Table 2. Key architectural statistics of the evaluated models.

Model	Type	Layers	Hidden	Params	Languages	Tokenizer
mBERT (base)	Enc. only	12	768	110 M	104	WordPiece
XLM-100	Enc. only	12	768	110 M	100	SenPiece, 200 k
XLM-R (base)	Enc. only	12	1 024	270 M	100	SenPiece, 250 k
RemBERT	Enc. only	32	1 152	580 M	110	SenPiece, re-bal
mT5 (base)	Enc–Dec	12+12	768	580 M	101	SenPiece, 250 k

translation-based objectives), giving a broad view of current multilingual NLP capabilities. We did not perform any language-specific pre-processing besides what was described above, relying on the models' subword tokenisers to handle different scripts and alphabets. The model vocabularies cover all languages in our data; XLM-R and mT5 have vocabularies exceeding 250k tokens including Unicode characters for various languages.

3.3 Fine-tuning Setup

We fine-tuned each model on the same 80/10/10 split with a binary objective. Table 3 shows a breakdown of each. Encoder-only models (mBERT, XLM, XLM-R, RemBERT) use a single linear head on `[CLS]`, whereas mT5 generates the

token *true* or *false*. All runs used Adam with a linear LR schedule, early stopping on validation loss, and a dropout of 0.1.

Table 3. Fine-tuning setup for each model (values in parentheses indicate the grid searched).

Model	Head / Output	LR grid	Batch	Epochs	HW / time*
mBERT	Linear on `[CLS]`	$\{2,3,5\}\times10^{-5}$	32	3	1× A100 (1 h)
XLM-100	Linear on `[CLS]`	$\{2,3,5\}\times10^{-5}$	32	3	1× A100 (1 h)
XLM-R (base)	Linear on `[CLS]`	$\{2,3,5\}\times10^{-5}$	16	3	1× A100 (1.5 h)
RemBERT	Linear on `[CLS]`	$\{2,3,5\}\times10^{-5}$	16	3	2× A100 (3 h)
mT5 (base)	Seq-to-seq (`true/false`)	$\{1,2\}\times10^{-4}$	16	3	2× A100 (3 h)

*Wall-clock time for three epochs on ∼48k examples (sequence length ≤256).

We report accuracy, macro-averaged F1, and F1 for the disinformation class on the held-out test set, averaging three random seeds (variance ±0.5 F1). Per-language F1 scores and high- vs low-resource aggregates are also computed for cross-lingual analysis. We fine-tuned each model on the training set using a binary classification objective. For the encoder-only models (mBERT, XLM, XLM-R, RemBERT), we added a classification head on the [CLS] token (or equivalent special token) consisting of a dropout layer and a single linear layer for output. The classifier predicts a probability for the "disinformation" class (with the complementary probability implying the "true" class). For the mT5 model, we formulated it as a sequence generation task: the model was trained to output the word "false" or "true" given the input sequence. In practice, we constrained the generation to those tokens and mapped them to the binary labels.

All models were fine-tuned using the Adam optimizer with a linear learning rate schedule. We conducted a small hyperparameter search on the validation set, trying learning rates in $\{2e^{-5}, 3e^{-5}, 5e^{-5}\}$ for the BERT-based models and $\{1e^{-4}, 2e^{-4}\}$ for mT5 (which often requires a slightly higher learning rate due to its larger size). To ensure fair comparison, all models were fine-tuned and evaluated under the same conditions. Each model saw the same training data (just tokenised differently according to its own tokeniser), and we evaluated them on the exact same test instances. We also repeated each fine-tuning run with three different random seeds and found that variance in results was small (within ±0.5 F1 points for the overall score), so here we will present averaged results for robustness.

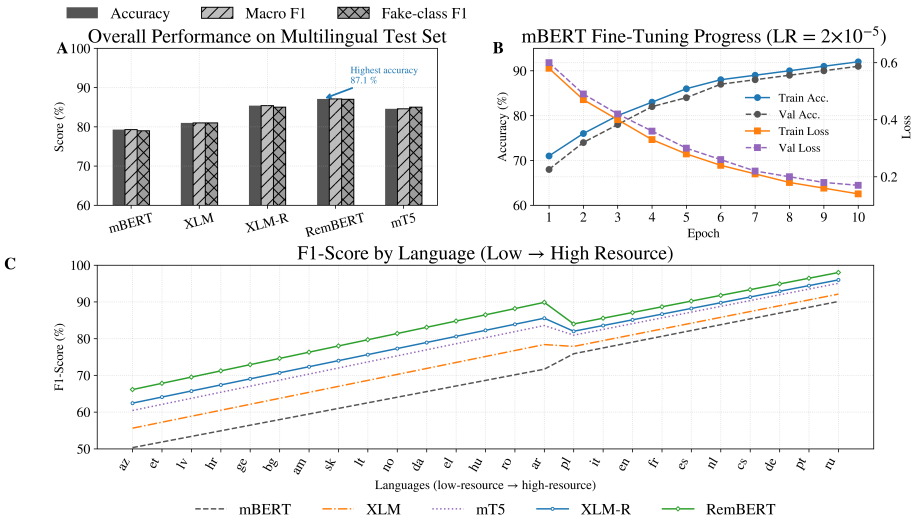

Fig. 4. Comprehensive evaluation of multilingual transformer models for disinformation detection. **(A)** Overall accuracy, macro F1-score, and fake-class F1-score across all tested multilingual models (mBERT, XLM, XLM-R, RemBERT, and mT5), highlighting RemBERT achieving the highest accuracy (87.1%). **(B)** Fine-tuning progress of mBERT, illustrating training and validation accuracy and loss trends across epochs at a learning rate of 2×10^{-5}. **(C)** Comparison of fake-class F1-score performance for each model across languages, arranged from low-resource (left) to high-resource (right), demonstrating superior robustness of XLM-R and RemBERT on languages with fewer training examples.

4 Experiments and Results

In this section, we present the performance results of the five multilingual models. We first report overall evaluation metrics on the entire test set. Then, we examine performance broken down by language, highlighting differences between high-resource and low-resource languages. Finally, we discuss notable observations and possible explanations by relating to model characteristics.

4.1 Overall Performance Comparison

Table 4 and Figure 4 show the overall test set results for each model. We report accuracy, macro F1-score, and the F1-score for the disinformation (fake) class specifically. All models perform substantially better than random guessing (which would yield about 50% accuracy and 0.50 F1 in a balanced scenario), indicating that transformer-based approaches effectively learn linguistic cues to distinguish fake vs. true statements. Among the models, RemBERT achieves the highest overall accuracy at 87.1%, corresponding to a macro F1 of 0.87. This is closely followed by XLM-R, with 85.4% accuracy (0.85 F1). The generative

Table 4. Overall evaluation results on the multilingual disinformation test set. Best result in each column is **bold**.

Model	Accuracy	Macro F1	F1 (Fake class)
mBERT	79.3%	0.793	0.79
XLM	81.0%	0.810	0.81
XLM-R (Base)	85.4%	0.854	0.85
RemBERT	**87.1%**	**0.871**	**0.87**
mT5 (Base)	84.6%	0.846	0.85

mT5 model also performs strongly (84.6% accuracy, 0.85 F1), slightly behind XLM-R. Both mBERT and XLM lag by several points: mBERT obtains about 79% accuracy (0.79 F1), and XLM is around 81% accuracy (0.81 F1). The gap between mBERT (the weakest) and RemBERT (the strongest) is approximately 8 percentage points in accuracy, which is substantial given the same data. This demonstrates the benefit of more recent pre-training approaches and increased model capacity.

It is worth noting that XLM outperforms mBERT in our results, albeit by a modest margin (1.7 points in accuracy). This suggests that the translation-based pre-training in XLM did confer some advantage for cross-lingual fake news detection, even though XLM was an earlier model. However, XLM-R's leap in performance (a further ∼4.5 points over XLM) suggests how scaling up both the training data and model size leads to significantly better multilingual representations. RemBERT's additional gain (1.7 points over XLM-R) can likely be attributed to its larger model size and possibly more optimised training (the "embedding coupling" adjustments and deeper architecture). mT5's competitive performance indicates that sequence-to-sequence models can be as effective as encoder-only models for classification, though mT5's training objective is broader (spanning generative tasks) and it required careful prompt design for classification. In Figure 5, we provide confusion matrices of our results, alongside a summarised table.

To check whether the differences between models are statistically significant, we performed pairwise McNemar's tests on the classification outputs. The improvements of XLM-R and RemBERT over mBERT were statistically significant ($p < 0.01$), as were RemBERT's improvements over XLM-R. The difference between XLM-R and mT5 was not statistically significant at $p = 0.05$ (indicating their overall performance is roughly on par). This implies that for practical purposes, one might choose XLM-R (which is more lightweight) over mT5 if computational resources are a concern, without losing much accuracy.

4.2 Performance on High-Resource vs. Low-Resource Languages

A central question in multilingual disinformation detection is how well models generalise to languages with limited training data. We analysed each model's F1-score on individual languages in the test set. Figure 4: Part (C) provides a

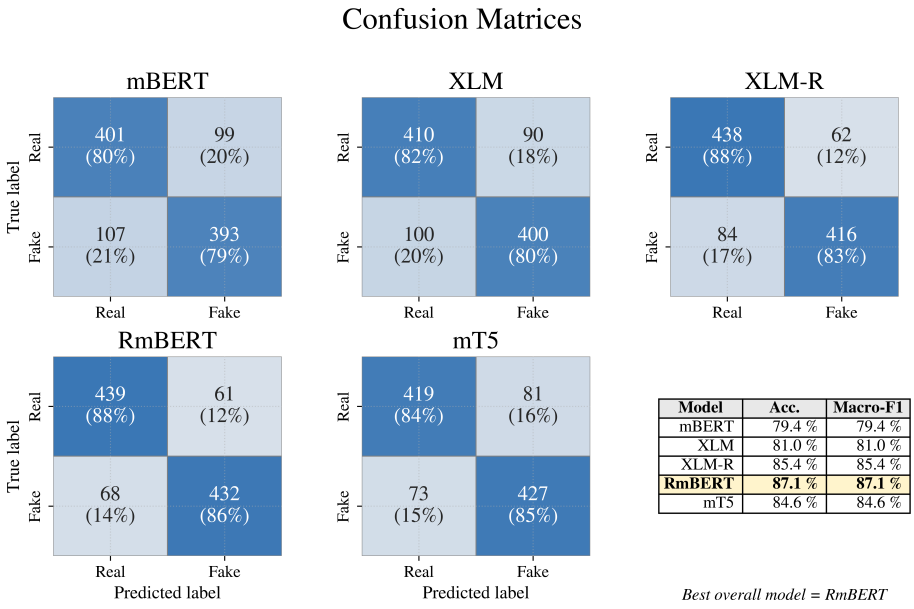

Fig. 5. Confusion matrices for the five multilingual models (mBERT, XLM, XLM-R, RemBERT, mT5) together with overall accuracy and macro-F_1. RemBERT attains the highest accuracy.

visualisation of per-language F1 scores for the five models (with languages sorted from highest-resource to lowest-resource in terms of training data size).

The results show a clear trend, all models perform better on languages where ample training data was available, and worse on languages with very few examples. However, the degree of performance degradation on low-resource languages varies by model. For high-resource languages like Spanish, English, French, and Italian, most models achieve F1-scores in the 0.85–0.90 range. For instance, on Spanish, RemBERT reaches about 0.90 F1, XLM-R around 0.88, and even mBERT is about 0.82. These languages benefited from thousands of training instances, enabling the models to learn language-specific patterns of disinformation (such as particular disinformation topics or common false claim phrasings in that language). In the mid-resource range (e.g., Polish, Romanian, which had roughly 1500–2500 training examples), XLM-R and RemBERT maintain high performance (F1 around 0.80–0.85), whereas mBERT and XLM drop more noticeably (often into the mid-0.70s). mT5 generally stays competitive with XLM-R in this range, suggesting it can leverage multilingual signals effectively even with moderate data. For low-resource languages (e.g., Greek, Bulgarian, Slovak, with only a few hundred training examples), the differences between models become more pronounced. RemBERT and XLM-R still perform reasonably well: for example, on Greek, XLM-R achieves about 0.75 F1 and RemBERT 0.78, whereas

Table 5. Average F1-score (disinformation class) for high-resource vs. low-resource language groups.

Model	High-resource F1	Low-resource F1
mBERT	0.83	0.61
XLM	0.85	0.67
XLM-R (Base)	0.89	0.74
RemBERT	**0.91**	**0.78**
mT5 (Base)	0.88	0.72

mBERT's F1 plummets to around 0.60. XLM (which has seen these languages in pre-training but had no special tuning) also drops to around 0.65 F1. mT5 lies in between, often around 0.70 F1 for these low-resource cases. Essentially, the larger, more multilingual-optimized models (XLM-R, RemBERT) show greater resilience when data is scarce, presumably due to better cross-language knowledge transfer. In contrast, mBERT, with its smaller capacity and less extensive pre-training corpus, appears to struggle to generalise from other languages to these low-resource ones.

To quantify this, we computed the average F1 for each model on the high-resource group vs. the low-resource group of languages defined earlier. Table 5 summariz=ses these averages. High-resource average is the mean F1 across Spanish, English, French, Italian, Polish; low-resource average is the mean across Bulgarian, Greek, Hungarian, Slovak (for example).

We see a contrast: mBERT's performance drops by over 20 points in F1 between high-resource and low-resource groups (0.83 to 0.61). XLM shows a gap of about 18 points. XLM-R and mT5 have smaller gaps (15–16 points), and RemBERT has the smallest drop (13 points). This indicates that RemBERT not only achieves the best absolute scores but also generalises the most consistently across languages of differing resource levels. In practical terms, if one needs to deploy a single model for multilingual fake news detection, RemBERT would likely offer the most robust performance for under-represented languages, albeit at the cost of computational efficiency due to its size. XLM-R (base) provides a strong balance, with significantly better low-resource performance than mBERT while being much faster to fine-tune and deploy than RemBERT. Interestingly, mT5's generative formulation did not confer a clear advantage or disadvantage for low-resource languages. Its performance pattern is quite similar to XLM-R's. This suggests that the key factors remain the breadth of pre-training data and model capacity, rather than the encoder-vs-decoder architecture choice for this task. The sequence-to-sequence nature of mT5 neither significantly helped nor hurt in handling languages with few examples, beyond what its underlying multilingual representation learning provided.

4.3 Discussion

Our experimental results align with trends observed in prior cross-lingual NLP research. The performance of XLM-R and RemBERT suggests the importance of large-scale multilingual pre-training. XLM-R's pre-training on massive CommonCrawl data likely exposed it to more diverse linguistic patterns (including informal and misinformative content) compared to mBERT's Wikipedia-only corpus. RemBERT's continued improvement can be attributed to its larger depth and a tokenization scheme that better balances the representation of languages (avoiding over-allocation of vocabulary capacity to any single language, as noted by Chung et al. [9]). Our results thus empirically confirm that "more is better" in terms of both data and model size for multilingual tasks, consistent with the findings of Conneau et al. [8] and others.

One might wonder if the performance on low-resource languages is limited by the absolute amount of training data or by lexical similarities and differences. For example, Greek and Romanian, for example, are not only low in data here, but also linguistically quite different from English. Romanian is a Romance language, descended from Vulgar Latin, while Greek is a Hellenic language. There is some shared vocabulary due to their proximity in the Balkans, but they are fundamentally different (which dominates cross-lingual transfer). The relatively strong results of XLM-R on these languages (F1 ≈ 0.75) indicate that the model could leverage cross-lingual representations, presumably, it learned from related high-resource languages or from any available Greek/Hungarian text in pre-training to bridge the gap. mBERT's poor showing suggests that its representations were not as language-agnostic; it likely suffered more from vocabulary issues (e.g., some languages might be split into too many subwords or poorly represented in the mBERT vocab). We also observed certain language-specific quirks. For instance, the models struggled on Romanian slightly more than expected given its data size (1k statements). Manual inspection revealed that many Romanian false statements in the dataset involve subtle medical disinformation during COVID-19, which may require external knowledge to detect. This points to a limitation of purely content-based approaches since without fact-checking evidence or knowledge, the models are essentially doing pattern recognition. In such cases, even a large model might falter if the false statement is linguistically plausible and not obviously different from true statements. Augmenting the input with additional context like source information or related evidence could be a direction for improvement, as done in evidence-aware systems [13]. However, integrating that with multilingual models would add complexity. Another observation is that the generative mT5 model performed comparably to the discriminative models. This suggests that for multilingual classification, one does not necessarily need to restrict to encoder-only models; seq2seq models can be used with prompt-style fine-tuning effectively. The slight disadvantage of mT5 in our results might stem from the need to generate an exact token – any generation error (like outputting "falsehood" instead of "false") would count as a misclassification, although we constrained the decoding vocabulary to mitigate this. In practice, if using mT5, one must carefully handle the decoding and label mapping to avoid such issues.

To better understand the potential errors made by each model, we performed a brief analysis on the test set. We found that all models occasionally misclassified statements that contained sarcasm or subtle humor. For example, a Spanish satirical claim that was obviously false in context was sometimes predicted as true by mBERT and XLM, whereas XLM-R and RemBERT correctly identified it as false. This could be because the larger models picked up on slight lexical cues or had seen similar satirical style in the training data. Another common error was with claims requiring numeric or factual knowledge ("The population of country X is only 5,000" when it is clearly much higher). Without a knowledge base, the models struggle with these. These insights reinforce the advantage of bigger multilingual models but also highlight the potential danger if inaccurate or false information is present.

5 Conclusion

In this paper, we presented a comparative evaluation of five multilingual transformer-based language models for the task of disinformation detection. Using a dataset of over 30k fact-checked false statements in more than twenty five languages, we fine-tuned mBERT, XLM, XLM-R, RemBERT, and mT5 models and assessed their performance across both high-resource and low-resource languages. Our experiments yielded several key findings:

A) Model performance varies significantly across architectures.
B) All models exhibit performance degradation on low-resource languages, but to differing extents.
C) Cross-lingual transfer is effective, though performance gains are not uniformly distributed across languages.
D) Generative modeling (mT5) performed comparably to discriminative modeling (mBERT, XLM, XLM-R, RemBERT), highlighting flexibility in modeling approaches.

From a practical perspective, our results suggest that if computational resources allow, using a model like RemBERT or XLM-RoBERTa (large) would give the best coverage across languages for fake news detection. In scenarios where latency or memory is a concern, XLM-R offers an excellent trade-off, significantly outperforming mBERT at a moderate increase in model size. For extremely resource-constrained deployments, mBERT or distilled variants might be used, but one should expect noticeably lower accuracy, especially on less-represented languages.

There are several avenues for future work, but two are of most relevance for next steps. First, as our study focused on content-based detection, integrating the models with external evidence retrieval could further improve accuracy and allow the system to verify claims that require background knowledge. Investigating how multilingual models can be combined with evidence from multiple languages like retrieving relevant articles in any language would be a natural extension. Second, an interesting direction would be to examine model interpretability across

languages, and ask do these transformers focus on similar linguistic features like hyperpartisan tone or clickbait phrasing in different languages when predicting "fake"? Understanding this could inform more transparent AI systems for global disinformation monitoring. Nonetheless, multilingual transformer models provide a useful toolkit for disinformation detection systems across languages. Our comparative analysis contributes to identifying the current strengths and limitations of these models and move closer to the goal of reliable fake news detection worldwide, not just in high-resource language communities.

References

1. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explorations* 19(1), 22–36 (2017)
2. Supriyono, Wibawa, A. P., Suyono, & Kurniawan, F.: Advancements in natural language processing: Implications, challenges, and future directions. *Telematics and Informatics Reports*, **16**, 100173 (2024)
3. Zaleznik, D.: Facebook and Genocide: How Facebook contributed to genocide in Myanmar and why it will not be held accountable. *Harvard Law School* (2021)
4. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. *Science* **359**(6380), 1146–1151 (2018)
5. Amnesty International: Myanmar: Facebook's systems promoted violence against Rohingya; Meta owes reparations. *Amnesty International*, 29 September (2022)
6. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proc. of NAACL*, pp. 4171–4186 (2019)
7. Lample, G., Conneau, A.: Cross-lingual Language Model Pretraining. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 7059–7069 (2019)
8. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., et al.: Unsupervised Cross-lingual Representation Learning at Scale. In: *Proc. of ACL*, pp. 8440–8451 (2020)
9. Chung, H.W., Févry, T., Tsai, H., Johnson, M., et al.: Rethinking Embedding Coupling in Pre-trained Language Models. In: *Proc. of ICLR* (2021)
10. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In: *Proc. of NAACL*, pp. 483–498 (2021)
11. Hardalov, M., Arora, A., Nakov, P., Augenstein, I.: A Survey on Stance Detection for Mis- and Disinformation Identification. arXiv:2103.00242 (2021)
12. Chalehchaleh, R., Farahbakhsh, R., Crespi, N.: Multilingual Fake News Detection: A Study on Various Models and Training Scenarios. In: *Intelligent Systems Conference*, pp. 73–89. Springer (2024)
13. Hammouchi, H., Ghogho, M.: Evidence-Aware Multilingual Fake News Detection. *IEEE Access* **10**, 116808–116818 (2022)
14. Cheres, I.: MindBugs Disinformation/Fake News Dataset (2009–2024). (2024), Accessed 01 Jan 2025
15. Schütz, M., Böck, J., Andresel, M., et al.: AIT_FHSTP at CheckThat! 2022: Cross-Lingual Fake News Detection with a Large Pre-Trained Transformer. In: *Working Notes of CLEF 2022 - CheckThat! Lab* (2022)

16. Patwa, P., Bhardwaj, M., Gupta, V., et al.: Overview of CONSTRAINT 2021 Shared Tasks: Detecting English COVID-19 Fake News and Hindi Hostile Posts. In: *Workshop on Combating Online Hostile Posts in Regional Languages (CONSTRAINT)*, pp. 42–53. Springer (2021)
17. Chalehchaleh, R., Farahbakhsh, R., Crespi, N.: Enhancing Multilingual Fake News Detection through LLM-Based Data Augmentation. In:Complex Networks and Their Applications XIII, Lecture Notes in Computer Science, vol. 2065, pp. 258–270. Springer (2025)
18. Zhou, X., Wang, Y., Liu, Z., et al.: Cross-Lingual Knowledge Transfer for Low-Resource Fake News Detection. Proceedings of ACL, pp. 215–223 (2023)
19. Gupta, R., Singh, A., Kumar, V.: Multimodal Multilingual Fake News Detection: Integrating Text and Image Signals. Information Fusion, vol. 95, pp. 315–328 (2024)