Fake News Detection in Pakistan: A Multilingual Approach

Fatima Tariq
Habib University
Karachi, Pakistan
Email: ft07200@st.habib.edu.pk

Raza Hashim Nizamani
Habib University
Karachi, Pakistan
Email: rn07380@st.habib.edu.pk

***Abstract*—The rise of digital media has exacerbated the spread of misinformation, significantly undermining public trust, particularly in countries like Pakistan, where news is widely consumed in both English and Urdu. This paper presents a system designed to detect fake news across these two languages, addressing the unique linguistic and cultural challenges posed by Pakistan's media environment. The project aims to develop a classification model capable of categorizing news articles as True or False, considering the complex nature of misinformation and linguistic differences. The ultimate goal is to enhance media reliability and support fact-checking efforts in a multilingual context. A bilingual dataset combining 12,000 English and 900 Urdu samples was used to evaluate the system. The study highlights the challenges of dataset imbalances in both language and class distribution. Among the evaluated models, XLM-RoBERTa achieved the best accuracy (96.74%), showcasing the potential of multilingual pre-trained models for handling low-resource languages like Urdu.**

***Index Terms*—Fake news, misinformation, natural language processing, Urdu, English, bilingual classification, news verification, transformers, BERT, roBERTa**

I. Introduction

The proliferation of digital media has amplified the spread of misinformation, posing significant challenges to public trust in the media. This issue is particularly prominent in multilingual countries like Pakistan, where news consumption occurs in both English and Urdu. Inaccurate information can severely influence public opinion, sway political decisions, and destabilize social dynamics, making the need for reliable news verification tools more critical than ever.

In response to this growing issue, our project seeks to develop a fake news detection system tailored to the unique linguistic and cultural landscape of Pakistan. The system will classify news articles into two categories: True or False. The primary objective is to create a model that can work effectively across both English and Urdu, two languages that differ significantly in terms of grammar, vocabulary, and writing systems.

Our approach leverages advancements in natural language processing (NLP) to design a bilingual classification model capable of detecting misinformation in a way that accounts for the linguistic nuances and challenges of processing news in these two languages. This paper outlines the methodology used to develop this system, reviews recent advancements in the field of NLP for fake news detection, and discusses the performance metrics that demonstrate the effectiveness of our model.

II. Related Work

The detection of fake news has garnered significant attention in recent years, with various studies exploring innovative techniques across diverse languages and contexts. This section reviews key contributions that align with our goal of developing a bilingual fake news detection system for English and Urdu.

Shanmugavadivel et al. [1] focused on fake news detection in Dravidian languages using traditional machine learning algorithms, achieving a highest accuracy of 77.76% with Naive Bayes and Support Vector Machines (SVM). Their work established a strong baseline but emphasized the need for advanced deep learning models to handle linguistic complexities effectively.

In a study on Russian fake news detection, Kuzmin et al. [2] utilized linguistic features, discourse structures, and transformer-based embeddings like RuBERT. While RuBERT excelled in multiclass classification with an F1-score of 0.822, n-gram-based methods demonstrated superior performance for binary tasks, highlighting the enduring relevance of traditional approaches in certain scenarios.

Efforts in low-resource languages have also gained traction. Santosh et al. [3] explored Urdu fake news detection using machine learning classifiers, achieving 91% accuracy with XGBoost on a dataset of 900 samples. This study underscored the limitations of small datasets and the potential benefits of incorporating deep learning models for better generalization.

Addressing data scarcity, Harris et al. [4] introduced a benchmark dataset for Urdu fake news with over 10,000 samples. Their ensemble model, combining mBERT, XLNet, and XLM-RoBERTa, achieved 95.6% accuracy, demonstrating the power of multilingual pre-trained models for low-resource languages.

Large language models (LLMs) have also been investigated for fake news detection. Hu et al. [5] evaluated GPT-3.5 for Chinese and English datasets, proposing the ARG model to integrate LLM-generated rationales with fine-tuned small language models (SLMs). While SLMs outperformed LLMs individually, their combination achieved state-of-the-art per-

formance, demonstrating the complementary strengths of these models.

Lastly, Su et al. [6] addressed the challenges posed by mixed human- and machine-generated fake news. Using fine-tuned transformer models like RoBERTa, they observed robust performance across various content production stages, emphasizing the need for nuanced approaches to detect hybrid and machine-generated content.

These studies collectively highlight the potential of leveraging pre-trained models and cross-lingual techniques for fake news detection, particularly in linguistically diverse and resource-constrained settings. Our work builds on these advancements by introducing a bilingual system that addresses the unique challenges of detecting fake news in English and Urdu, thereby contributing to the broader goal of multilingual misinformation mitigation.

III. Methodology

A. Data Description

The aim of our approach is to be linguistically versatile, meaning we want it to be able to detect the authenticity in cases when the input is in Urdu or English. The reason for adding Urdu is that it is the predominant language for Pakistani news media, both print, television and social media. To implement it we created a combined two different datasets, one in English and one in Urdu, to train our model(s) on. Their individual details are listed below:

- **English Dataset**: This dataset has about **12,000 entries** related to Pakistani news in English only. The data for it was gathered from the web sources by searching for keywords such as Pakistan, Pakistani etc.. All entries as categorized as one of the following labels; True, Partly True, False, Fake, Doctored, Hoax, Misleading, Mixture, Half True, and Satire. For this categorization, the authors of dataset relied on tools such as Google Fact Checker and Politifact. These services are some of the most reknowed and reliable for this sort of classification. Even though the original dataset consisted of multiple labels to simplify our classifications and to make the two datasets compatible we limited ourself to just **True** and **False** labels. The original breakdown of classes was as follows in TABLE I;

TABLE I
Original Class Breakup

Label	Count
True	10461
Partly True	3
False	1711
Fake	11
Doctored	3
Hoax	1
Misleading	90
Mixture	1
Half True	62
Satire	5

For simplification we converted the labels True, Half True and Partly True to the single label True and converted the labels False, Fake, Hoax, Doctored, Misleading, Mixture, and Satire to False. After doing so we were left with the following dataset shown in TABLE II;

TABLE II
Post Simplification

Label	Count
True	10526
False	1822

- **Urdu Dataset**: The dataset we relied on for Urdu entries was on the other hand much smaller. In total, it consisted of only **900 entries**. These entries were sourced from various different sources such as BBC Urdu News, CNN Urdu, Express-News, Jung News, Noway Waqat, and others. The dataset contains entries regarding news from 5 different domains; Sports, Health, Technology, Entertainment, and Business. Doing so allows it to be diverse in terms of both content and style. As for the classification in the dataset, it is much simpler than the English one with just two labels; **Real** or **Fake**. The class breakdown was is shown in TABLE III;

TABLE III
Urdu Dataset

Label	Count
Real	500
Fake	400

The structure in which the data was stored was as follows; there was a primary folder "Urdu Fake News Dataset" consisting of sub-folders "Real" and "Fake". Each of these folders then further had two sub-folders "Train" and "Test". These folders than consisted of individual text files for each entry. However, for our approach, we combined all this data into one data frame to be later combined with the English one.
- **Combining the two Datasets**: To combine the two datasets we first added a 'source' column to each dataset to determine whether the entries originate from the English Dataset or the Urdu one. Once that was done and the labels for the English dataset were simplified, we simply concatenated the two data frames, more specifically the columns, 'text', 'cleaned labels' (simplified and formatted to match each other) and 'source' (to be able to determine original dataset).

B. Word Frequency Analysis

As part of our analysis, we performed a word frequency analysis on both the English and Urdu datasets. This allowed us to identify common words and themes in the fake news articles, which can help in understanding the patterns of misinformation. The analysis was conducted separately for each language, ensuring that linguistic differences were properly

accounted for. For each dataset, we focused on the "Fake" news category, as this is the primary target for detection in the model.

1) English Fake News Dataset: The English dataset contained approximately 12,000 entries and consisted of news articles in English that were labeled as either "True" or "False." We conducted a frequency analysis on the "Fake" news articles to identify the most common non-stopwords. This step was crucial as it allowed us to pinpoint recurring topics or narratives that are often found in misleading content. The results showed that certain words related to political, social, and economic events were highly frequent, such as "trump," "president," "Pakistan," and "states," reflecting the themes typically found in fake news about Pakistani affairs.

To further refine the analysis, stopwords—common words like "the," "is," "at," etc.—were removed from the dataset, as they often do not contribute much to the understanding of the content's subject. After cleaning the data, the most frequent words were plotted to visualize their distribution. This frequency distribution shown in Fig. 1 provides valuable insights into the topics that dominate the fake news landscape in English about Pakistan.

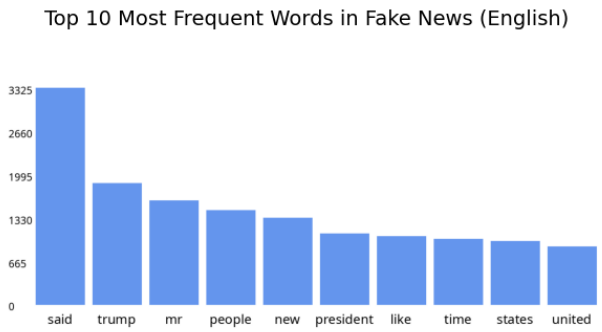

Fig. 1. Top 10 Most Frequent Words in Fake News (English)

2) Urdu Fake News Dataset: Similar to the English dataset, the frequency analysis for the Urdu "Fake" news articles focused on identifying the most common words, excluding stopwords. The stopwords were sourced through multiple datasets found online. After processing, a bar chart was generated to visualize the top 20 most frequent words in the Urdu fake news articles. Among these, "pakistan," "company," "tehqeeq," appeared among the most frequent, indicating that these words are central to many fake news narratives. The frequency distribution is shown in Fig. 2.

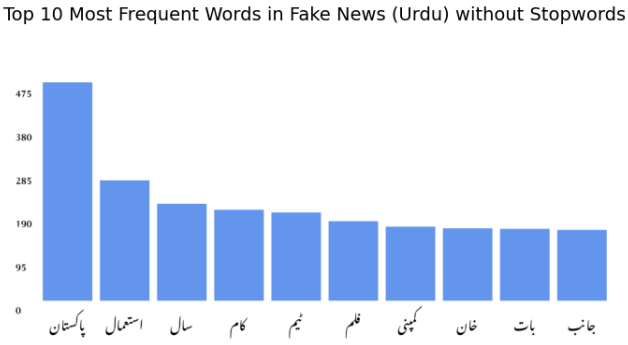

Fig. 2. Top 10 Most Frequent Words in Fake News (Urdu)

It is important to note that, due to the relatively small size of the Urdu dataset, the frequency analysis may not reflect the full diversity of topics that could appear in a larger dataset. The smaller sample size may lead to an overrepresentation of certain words or topics, particularly those that are more common in the collected sources.

C. Experimentation

*1) **Models**:* In this study, we experimented with three different models for the task of Fake News Detection, aiming to compare and contrast their performance on a multilingual language dataset. The first two models were BERT and RoBERTa. Both of these models were fine-tuned on English-language data for the Fake News Detection task, allowing us to assess their ability to generalize to a different, multilingual dataset for the same experiment during our experiments. The third model we evaluated was XLM-RoBERTa, a multilingual variant of RoBERTa, which is specifically designed to work across multiple languages(including Urdu). For XLM-RoBERTa, we used the vanilla (pre-trained) version without any further fine-tuning, providing an additional baseline to evaluate the benefits of multilingual capabilities compared to the monolingual models.

For the baseline, we utilized Logistic Regression, a simple yet effective traditional machine learning model, to establish a benchmark for comparison. This allowed us to assess whether the transformer models could significantly outperform a conventional approach, which might still be viable for smaller or less complex datasets.

*2) **Parameters**:* To ensure a fair comparison across the models we kept the parameters mostly the same across their training. We constrained our training to 3 epochs, as we saw that this was sufficient for the models to reach convergence and achieve saturation, beyond which performance did not significantly improve. This decision was based on both saving training time and computation and the observation that performance gains after 3 epochs were marginal.

We performed training in batches of 32 samples, which we found provided a good trade-off between training speed and model performance. A batch size of 32 is a commonly used value in transformer-based models, balancing computational efficiency with model generalization.

For the learning rate, we selected a value of 2e-5, which is a typical choice for fine-tuning pre-trained transformer models such as BERT and RoBERTa. This learning rate has been shown to provide stable convergence and prevent overfitting, especially for large models like these.

In terms of optimization, we used the AdamW optimizer, a variant of Adam that incorporates weight decay regularization. AdamW is widely used in transformer training due to its ability to effectively handle sparse gradients and its success in preventing overfitting, particularly in models with large parameter spaces like BERT, RoBERTa, and XLM-RoBERTa.

IV. Results and Discussion

In this section, we present the outcomes of our experiments comparing three different models for the task of Fake News Detection: BERT, RoBERTa, and XLM-RoBERTa, along with a baseline Logistic Regression model. We evaluate the performance of these models based on their ability to detect fake news in a multilingual context. We will be focusing on key metrics such as accuracy, Precision, Recall, and F1-score for our analysis. The results of the experiments provide insights into the effectiveness of transformer-based models, including the impact of multilingual pre-training, and offer a comparison against a traditional machine learning baseline.

A. BERT Model

The BERT model was the lowest performer out of our three main approaches. It yielded the second-best accuracy of **95.55%** on the combined dataset was the lowest in all other metrics. Despite being third the results it shows are still more than satisfactory.

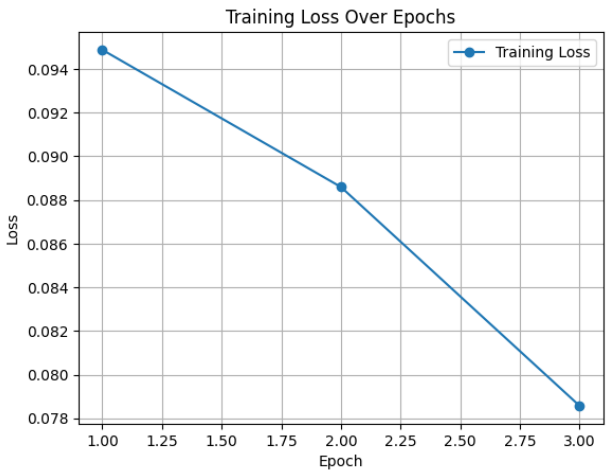

Fig. 3. BERT Loss across epochs

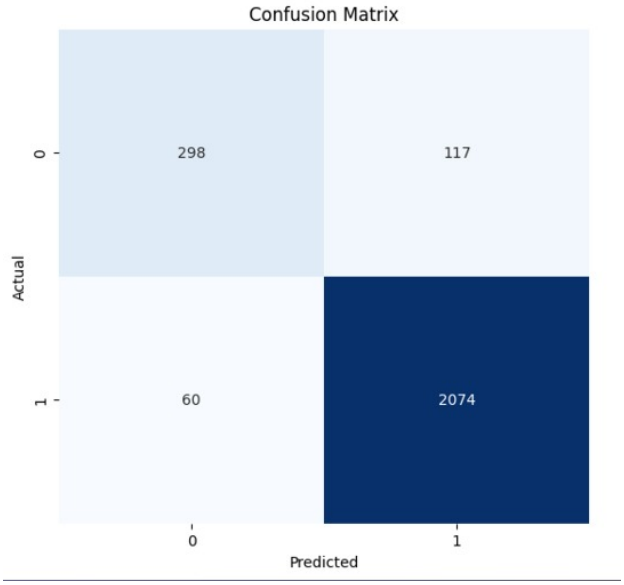

Fig. 4. BERT Confusion Matrix

B. RoBERTa Model

The RoBERTa model had the lowest Accuracy of **95.24%** but was the best performer in the rest of the metrics with the highest F1-score of **0.9497**.

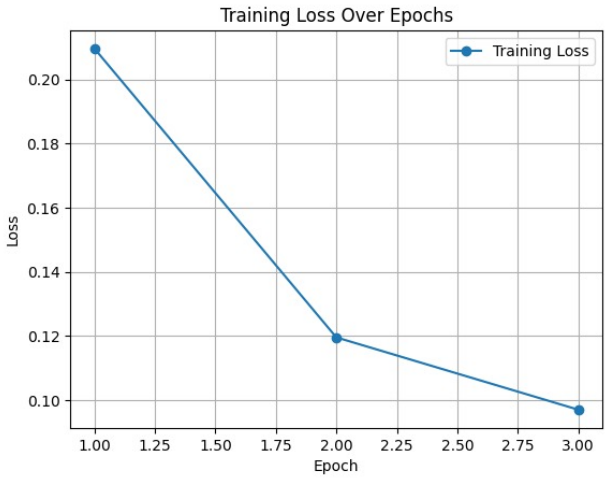

Fig. 5. RoBERTa Loss across epochs

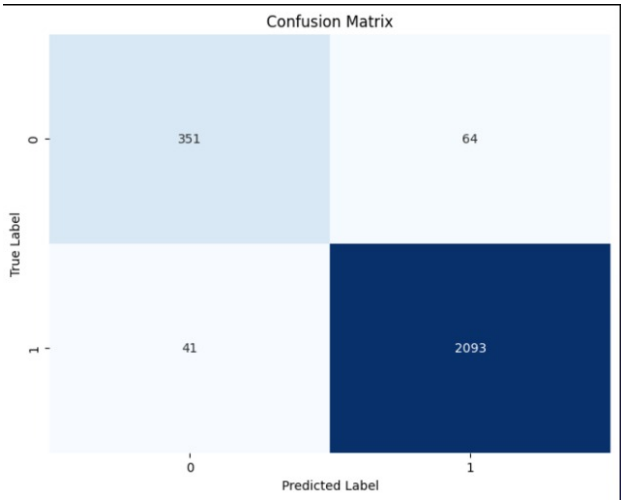

Fig. 6. RoBERTa Confusion Matrix

C. XLM-RoBERTa

The XLM-RoBERTa model, that was pre-trained on a multilingual corpurs which included Urdu had the best accuracy of **96.74%**. It also managed to be the second best in the all other metrics not to far off from the RoBERTa model. This aligns with the existing literature as XLM-RoBERTa was usually the best performer in multilingual tasks.

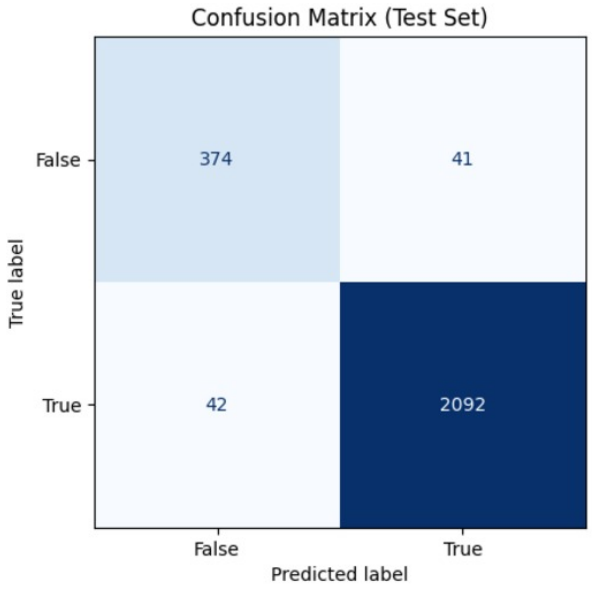

Fig. 7. XLM-RoBERTa Confusion Matrix

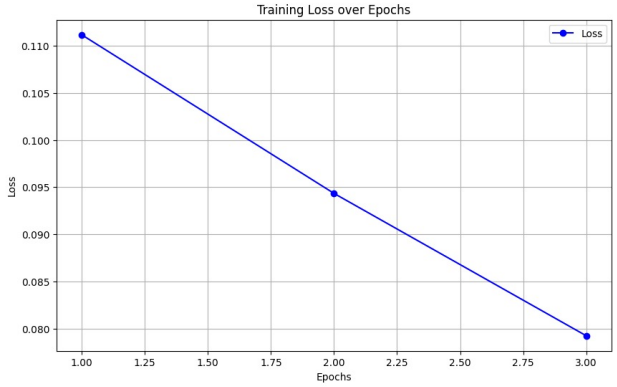

Fig. 8. XLM-RoBERTa Loss across epochs

D. Discussion

The overall results of all our approaches are shared below in TABLE IV

TABLE IV
OVERALL RESULTS

Metric	BERT	RoBERTa	XLM-RoBERTa	Logistic Regression
Accuracy	95.88%	95.24%	96.74%	89.0%
Precision	0.9329	0.9550	0.9399	0.88
Recall	0.9132	0.9524	0.9407	0.90
F1-Score	0.9227	0.9497	0.9403	0.88

The results indicate that there is very little difference between the three different approaches, with all models achieving high accuracy and performance. This suggests that all three models are well-suited for the Fake News Detection task, but several factors could explain why the performance is so close.

Firstly, the BERT and RoBERTa models were fine-tuned specifically for the Fake News Detection task, which could explain their strong performance. Fine-tuning allows these models to adapt more effectively to the specific nuances and patterns present in the fake news dataset, leveraging their pre-trained knowledge of the task. This pre-training gives them a significant advantage in understanding context, identifying relationships between words, and distinguishing between real and fake news, which are crucial aspects of the task.

On the other hand, the third model, XLM-RoBERTa, although not fine-tuned for the specific task, has the added benefit of being a multilingual model. It was pre-trained on a much larger and diverse dataset, which included a wide variety of languages, including Urdu. This multilingual pre-training enables the model to generalize better across languages and detect patterns in the Urdu text present in the fake news dataset. The fact that this model performs similarly to the others, despite not being fine-tuned specifically for Fake News Detection, highlights the power of multilingual models in handling diverse linguistic data.

Moreover, the results being so high, even with relatively limited training (of just 3 epochs) emphasize a critical observation, that the task of Fake News Detection is likely reaching a point of saturation. The small differences in

performance between the models suggest that the margins for improvement are becoming very narrow. As the models' accuracy approaches near-perfect levels, achieving significant gains becomes increasingly difficult, and even small changes in architecture or training strategies may have minimal impact. This is a common challenge in mature domains where models are already optimized and only incremental improvements remain. The high baseline performance also emphasizes that these models are robust enough to handle the complexity of Fake News Detection.

While the classwise breakdown of the experiments as shown in TABLE V reveals that the models perform significantly better for the True class (Class 1) than for the Fake class (Class 0), this observation is primarily attributed to the substantial class imbalance present in the dataset. The Dataset exhibits a nearly 17:3 ratio of True to Fake news instances, which means that the models are exposed to far more examples of the True class during training. This naturally leads to better performance for the True class as the model becomes more familiar with its patterns.

TABLE V
CLASS-WISE BREAKDOWN

Model	Metric	Fake	True
BERT	Precision	0.88	0.96
	Recall	0.78	0.98
	F1-Score	0.82	0.97
RoBERTa	Precision	0.895	0.970
	Recall	0.846	0.981
	F1-Score	0.870	0.976
XLM-RoBERTa	Precision	0.90	0.98
	Recall	0.90	0.98
	F1-Score	0.90	0.98
Logistic Regression	Precision	0.75	0.91
	Recall	0.55	0.97
	F1-Score	0.64	0.94

V. CONCLUSION AND FUTURE WORK

This study presented a multilingual fake news detection system designed for the unique linguistic and cultural context of Pakistan, addressing the challenges posed by news dissemination in both English and Urdu. By leveraging advanced transformer-based models, including BERT, RoBERTa, and XLM-RoBERTa, the system demonstrated high accuracy, with XLM-RoBERTa achieving the best performance (96.74%) due to its multilingual pre-training.

The results underscore the effectiveness of transformer architectures for multilingual tasks while highlighting the potential of pre-trained models in handling low-resource languages like Urdu. However, the findings also suggest that performance improvements in fake news detection may be reaching a saturation point, given the narrow performance margins between models.

The study faced several limitations. A significant language imbalance was present in the dataset, with only 900 Urdu samples compared to 12,000 English samples, representing a 1:13 ratio. This disparity likely affects the model's ability to generalize across both languages, particularly in Urdu, limiting its practical applicability in real-world scenarios. Additionally, the dataset exhibited a class imbalance, with 11,000 samples labeled as True and only 2,200 as Fake. This imbalance impacted model performance, resulting in a 10-point difference in F1-scores between the two classes, indicating a bias toward identifying True news more effectively than Fake news.

Future work should focus on addressing these limitations by expanding the Urdu dataset to achieve better language and class balance, ensuring a more robust and equitable performance across languages. Pre-training models specifically on Urdu and exploring ensemble methods that leverage the strengths of multiple architectures could enhance the system's capabilities further. Additionally, revisiting the multi-class framework could provide a more detailed understanding of different types of misinformation, enabling the system to address diverse real-world challenges more effectively.

REFERENCES

[1] K. Shanmugavadivel, M.Subramanian,"BeyondTech@DravidianLangTech2024: Fake News Detection in Dravidian Languages Using Machine Learning," in *Proc. 4th Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, St. Julian's, Malta, 2024, pp. 124–128. https://aclanthology.org/2024.dravidianlangtech-1.20

[2] G. Kuzmin, D. Larionov, D. Pisarevskaya, and I. Smirnov, "Fake news detection for the Russian language," in *Proc. 3rd Int. Workshop on Rumours and Deception in Social Media (RDSM)*, Barcelona, Spain, 2020, pp. 45–57. https://aclanthology.org/2020.rdsm-1.5

[3] Santosh Na, Parkash Na, Toto Na, and Dr. Zarmeen Nasim. 2022. Verify: Breakthrough accuracy in the Urdu fake news detection using Text classification. In Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation, pages 740–748, Manila, Philippines. Association for Computational Linguistics. https://aclanthology.org/2022.paclic-1.81

[4] S. Harris, J. Liu, H. J. Hadi, and Y. Cao, "Ax-to-Grind Urdu: Benchmark Dataset for Urdu Fake News Detection," in *Proceedings of the 2023 IEEE 22nd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 2024, pp. 2440–2447. https://doi.ieeecomputersociety.org/10.1109/TrustCom60117.2023.00343

[5] B. Hu, Q. Sheng, J. Cao, Y. Shi, Y. Li, D. Wang, and P. Qi, "Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection," in Proceedings of the AAAI Conference on Artificial Intelligence, 2023. https://doi.org/10.48550/arXiv.2309.12247

[6] J. Su, C. Cardie, and P. Nakov, "Adapting Fake News Detection to the Era of Large Language Models," in Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, 2024, pp. 1473–1490. https://doi.org/10.18653/v1/2024.findings-naacl.95