Detection of Human and Machine-Authored Fake News in Urdu

Anonymous ACL submission

Abstract

The rise of social media has amplified the spread of fake news, now further complicated by large language models (LLMs) like ChatGPT, which ease the generation of highly convincing, error-free misinformation, making it increasingly challenging for the public to discern truth from falsehood. Traditional fake news detection methods relying on linguistic cues also become less effective. Moreover, current detectors primarily focus on binary classification and English texts, often overlooking the distinction between machine-generated true vs. fake news and the detection in low-resource languages. To this end, we updated the detection schema to include machine-generated news with a focus on the Urdu language. We further propose a hierarchical detection strategy to improve the accuracy and robustness. Experiments show its effectiveness across four datasets in various settings. We release our collected datasets and code in URL withheld.

1 Introduction

Fake news detection aims to identify false or misleading information presented in news (Shu et al., 2019). With the rise of unrestricted social media, users can post virtually anything, accelerating the spread of misleading information. A substantial percentage of content shared on social media is found to be fake, making it a challenge for the general public to distinguish truth from falsehood. A recent study revealed that 48% of individuals across 27 countries have been misled by fake news, believing a false story to be true before later discovering it is fabricated.[1] This phenomenon may have serious consequences, including influencing public opinion, undermining democratic processes, and exacerbating societal divisions (Tandoc Jr et al., 2018; Lewandowsky et al., 2017). Effective fake news detection is thus crucial for maintaining a reliable society and ensuring the integrity of information.

While many studies exist for English fake news detection, research on low-resource languages such as Urdu remains under-explored (Ahmed et al., 2017; Previti et al., 2020). Previous work treats fake news detection as a binary classification task, relying on linguistic features. However, the ease of access to LLMs like GPT-4o (OpenAI, 2024) now enables propagandists to produce endless content mimicking journalistic tone with minimal errors, greatly complicating the task of evaluating the veracity of any given text (Wang et al., 2024b,a). Additionally, LLMs are increasingly utilized by journalists and media organizations, thereby blurring features that might help distinguish fake and real news. Currently, the publicly available Urdu datasets consist solely of human-written text (Amjad et al., 2020b; Akhter et al., 2021), which limits the amount of reliable data for developing new effective detection methods.

To fill this gap, we collected machine-generated news based on four existing datasets, each spanning short news subtitles and long articles. These datasets resulted in four four-label datasets comprising *human fake*, *human true*, *machine fake*, and *machine true* categories, as previously done by Su et al. (2023a). Transformation of the binary problem into four labels improves robustness against machine-generated fake news. It also enables nuanced analysis to distinguish human- from machine-authored content, thereby improving detection accuracy and furthering the development of useful training datasets to explore the balance between human- and machine-written examples.We found that baseline detectors using fine-tuned RoBERTa are not robust, tending to mispredict *machine true* and *machine fake* to other classes.

To address this, we propose a hierarchical method that breaks down the original four-class

[1] https://redline.digital/fake-news-statistics/

problem into two subtasks: machine-generated text detection and fake news detection, as illustrated in Figure 2. Experiments demonstrate that the proposed approach outperforms (in terms of accuracy and F1-scores) the baseline across both tasks in-domain and cross-domain settings, suggesting effectiveness and robustness. Our contributions are summarized as follows:

- We collect the first Urdu dataset for machine-generated fake and true news.
- We propose an effective hierarchical approach for four-label fake news detection, that is more accurate and robust than fine-tuned RoBERTa on four labels.
- We conduct a detailed analysis investigating (1) reasons for low accuracy in cross-domain settings, and (2) the impact of data augmentation in machine-generated text detection task on enhancing overall fake news detection.

2 Related Works

This section reviews previous research on (1) methods for detecting fake news, with an emphasis on general approaches and those specific to Urdu, and (2) techniques for acquiring machine-generated real and fake news.

General Approach for Fake News Detection Fake news detectors vary in terms of input features and model architectures. Features involve content features, social features, temporal features, or combinations of these (Shu et al., 2017). Content features encompass details such as term frequency (Ahmed et al., 2017), sentiment (Bhutani et al., 2019), and parts of speech (Balwant, 2019). Social features are primarily used on social media platforms and include information such as friends' circles, pages followed, and reactions to posts (Sahoo and Gupta, 2021). Temporal features consist of time-related aspects that indicate when a given post was released. For example, Previti et al. (2020) proposes a Twitter-based fake news detector that incorporates time series data along with other features and reports favorable results.

Existing research explores various model architectures, ranging from simplistic machine learning (ML) algorithms to advanced transformers (Vaswani et al., 2017). For instance, Raza and Ding (2022) introduces an encoder-decoder transformer that leverages content and social data for early detection. Unsupervised methods aim to circumvent the labor-intensive task of labeling and utilize various heuristics in clustering. Yin et al. (2007) suggests that a website's credibility is linked to its consistency in providing accurate information. Similarly, Orlov and Litvak (2019) proposes a heuristic to indicate that coordinated propagandists tend to exhibit similar patterns.

Urdu Fake News Detection Research on the Urdu language is under-explored. Existing studies often exhibit a lack of diversity in the features and the model architectures. Kausar et al. (2020) employs n-grams and BERT embeddings as features, and logistic regression and CNNs as models for training the classifiers. However, translation versions of datasets do not necessarily reflect the actual news lexicon of the target language in real-world scenarios. Similarly, Amjad et al. (2020a) compares models trained on organically labeled Urdu fake news data against those trained on English fake news data translated into Urdu. It shows models trained on organically-labeled Urdu data outperform those trained on translations.

For Urdu datasets, Kausar et al. (2020) translates an English dataset Qprop (Barrón-Cedeno et al., 2019), into Urdu using Google Translate. Akhter et al. (2021) creates an Urdu fake news dataset by semi-automatic translation of an existing English fake news dataset, and uses ensemble approaches and content features for model training. In addition, three commonly used fake news datasets are specifically curated in Urdu: *bend-the-truth* (Amjad et al., 2020b), *ax-to-grind* (Harris et al., 2023), and *UFN2023* (Farooq et al., 2023). The work presented here uses these datasets for experiments that are detailed in Section 3.1.

Machine Generated Text What prompts have been used to generate paraphrased text via LLMs? Zellers et al. (2019) trains a model GROVER, which can generate and identify fabricated articles. Huang et al. (2022) uses BART for mask-infilling to replace salient sentences in articles with plausible but non-entailed text, ensuring disinformation through self-critical sequence training with an NLI component. Similarly, Mosallanezhad et al. (2021) proposes a deep reinforcement learning-based method for topic-preserving synthetic news generation, controlling the output of large pre-trained language models. All of these studies focus on generating fake news. However, LLMs are now utilized by news organizations and journalists, requiring a new schema of generating machine true news. Su et al. (2023b) presents a

Structured Mimicry Prompting approach for generating both machine fake and true news using *GPT-4o*, in which LLM understands the title and article body, and generates a similar text.

3 Dataset Collection

3.1 Datasets

Four publicly available Urdu fake news datasets are used to train models, with the creation of new data for two classes: *machine true* and *machine fake*. The datasets are as follows.

Dataset 1: *Ax-to-Grind Urdu* The latest Urdu fake news dataset published earlier this year is *Ax-to-Grind Urdu*. It has 10,083 samples related to fifteen different domains with approximately equal distribution of fake and true classes. Harris et al. (2023) maintains the originality of the corpus by keeping only the original news headlines. For real news headlines, data was collected from authentic news websites such as BBC Urdu, Jang, Dawn News, etc. Fake news headlines were collected from two of arguably the most controversial news websites: Vishwas News and Sachee Khabar. Additionally, some fake news was collected through crowd-sourcing. Professional journalists were hired to fact-check each individual news sample and label it accordingly.

Dataset 2: *UFN2023* This dataset was constructed using a hybrid approach that involved authentic news websites for real news and samples from the fake category of an English dataset translated (supervised) into Urdu fake news. Additionally, some obvious fake news headlines from *Vishwas News* were also included. The dataset contains 4,097 samples across nine different domains, such as health, sports, technology, and showbiz. Of these, 1,642 samples belong to the real news category, while 2,455 belong to the fake category.

Dataset 3: *UFN Augmented Corpus* *UFN Augmented Corpus* is another publicly available Urdu Fake News dataset. Akhter et al. (2021) randomly selected two thousand news articles from an English fake news dataset and translated them into another language using Google Translate with human supervision. The quality of the translations was manually verified, and articles, where the journalistic tone or meaning was lost, were replaced with different translated articles. The name of the English dataset has not been revealed in their work.

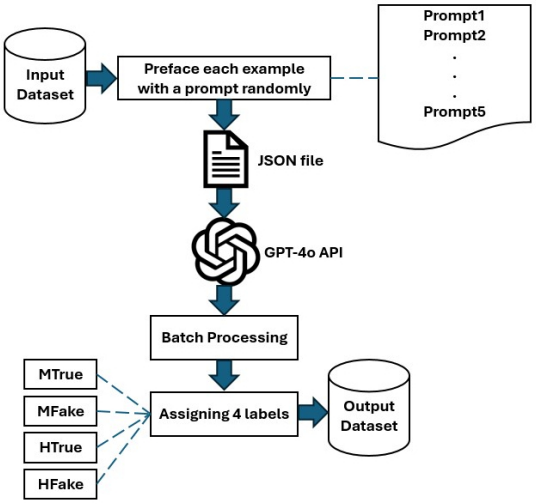


Figure 1: Machine generated News collection process

Out of two thousand translated articles, 968 news articles belong to the fake class and 1032 news articles belong to the true class.

Dataset 4: *Bend the Truth* This is perhaps one of the first publicly available Urdu fake news datasets; presented in Amjad et al. (2020b). It is relatively small, consisting of only 1300 (750 real, 550 fake) news articles, but the authors used a very interesting approach to keep the dataset organic. They collected true articles from authentic news websites and then hired journalists to rewrite them with a counterfactual narrative without losing the original journalistic tone.

Categorization Table 1 provides a summary of the four datasets used in this work. The text was tokenized using the word tokenizer from the NLTK library for Urdu[2], segmenting based on spaces. Token counts are carried out after stop words are removed. Based on the average counts in both categories, the datasets are classified as either Short or Long: Datasets 1 and 2 primarily contain short texts and news headlines, thus categorized as *Short*, whereas Datasets 3 and 4 consist of longer news articles, thus categorized as *Long*.

Split of Training, Development and Test Sets To create test sets for evaluation, a 20% random split was set aside for each of the four datasets before producing machine-generated text. This ensures a clear separation of the test set — both human-written and machine-generated data are entirely unseen during training. The validation set for

[2] https://www.nltk.org

Dataset	#Examples	#HF	#HT	#MF	#MT	$\hat{T}$(HF)	$\hat{T}$(HT)	$\hat{T}$(MF)	$\hat{T}$(MT)	Content	Category
Dataset1	10083	5053	5030	5053	5030	58.7	19.2	61.2	20.1	headlines	Short
Dataset2	4097	2455	1642	2455	1642	105.6	34.3	110.2	33.4	headlines	Short
Dataset3	2000	968	1032	968	1032	645.0	516.1	602.2	499.4	articles	Long
Dataset4	1300	550	750	550	750	134.1	198.0	101.3	211.6	articles	Long

Table 1: Statistical information of four datasets. Examples are organic samples. #= the number of news, $\hat{T}$=average tokens. **HF**: Human Fake, **HT**: Human True, **MF**: Machine Fake, **MT**: Machine True.

each experiment consists of 25% of the training set. Thus, 60% of the data is used for training, 20% for validation, and 20% for testing.

3.2 Machine-generated News Collection

GPT-4o was used to produce machine-generated news articles and short messages for both true and fake categories, paraphrasing original text using five different prompts. Figure 1 shows the overview of the generation and gold label assignment process. Each example is generated with one prompt randomly sampled from the five, using OpenAI batch generation. Afterward, gold labels are assigned. Labels of original articles are changed from True and Fake to *human true* and *human fake*. Machine-generated articles receive *machine true* and *machine fake* labels based on the labels of their parent news articles.

Generation Prompts Table 2 shows all five prompts used for generating machine data. We carefully designed and adjusted the prompts so that they can instruct GPT-4o to rephrase the provided article or headline, but keep the exact same meaning and stance without distortion.

Quality Control To ensure the data quality and avoid introducing factually false information, especially for *machine true* text, we randomly sampled 1008 examples from dataset 1 (10083 machine text) and asked three native Urdu speakers to review the machine-generated articles. They compared machine text with original articles and found that 9% samples have minor discrepancies from the human article, generally introducing additional context (i.e., more tokens).

Therefore, we filtered machine-generated cases where the number of tokens is different from the original articles or headlines by 20%, resulting in 712 problematic examples in dataset 1. We further analyzed these 712 examples and identified three problems: (i) 209 examples were not paraphrased; instead, GPT-4o responded with prompts like: *Please provide the news article for rephrasing.* (ii) in 403 examples, GPT-4o introduced information not present in the original text; and (iii) 100 paraphrased articles began with a preface from GPT-4o, such as: *Certainly! I can help you with rephrasing.*

To address these problems, all prompts were re-engineered by adding the following line before the last sentence: *Please directly rewrite without opening words like Of course I can help you with rewriting, and note that do not generate or extend extra information that is not included in the given article, DO NOT HALLUCINATE EXTRA INFORMATION.* These newly engineered prompts were applied to the problematic samples for dataset 1 and the remaining three datasets.

The generated text was re-evaluated using the methods discussed above, with particular emphasis on the *machine_true* class. Each generated article or headline was manually verified to ensure no narrative shift occurred compared to its corresponding *human_true* version, as any deviation would disqualify it from being labeled *machine_true*. Minor issues identified during this process were resolved by human annotators.

4 Methods

This section presents the baseline methods and our proposed hierarchical detection.

4.1 Baselines

We applied both traditional machine learning algorithms e.g., Support Vector Machine (SVM) with bag-of-words and TF-IDF features, and fine-tuning multilingual language models XLM-RoBERTa (XLM-R).

Linear SVM We first performed data cleaning (e.g. removing punctuation, stop words and URLs), followed by bag-of-words representation and TF-IDF feature extraction. Several machine learning models were trained, including SVM, Multinomial Naive Bayes, and Random Forest. We selected the

ID	Prompt
Prompt1	*I am going to provide you with an Urdu article. Please rewrite that article while keeping the same narrative. Feel free to completely change everything, every single word if you have to. In fact, I would appreciate it if there is very little similarity between the original article and what you write. Just the idea and narrative should essentially be the same. The article follows:*
Prompt2	*I will provide you with an Urdu article. Your task is to rewrite this article while maintaining the same core message and narrative. Ensure that the wording and structure are significantly different from the original. Here is the article:*
Prompt3	*Here is an Urdu article that I need you to rewrite. Please keep the underlying story and narrative intact, but rephrase it thoroughly so that it appears entirely new. Aim for minimal similarity to the original text. The article is as follows:*
Prompt4	*Please take the following Urdu article and rewrite it in such a way that the narrative and main idea remain unchanged, but the language and wording are entirely different. Your goal is to create a version with minimal resemblance to the original. Here is the article:*
Prompt5	*Given the following Urdu article, I need you to produce a rewritten version that preserves the same story and narrative. Feel free to alter the wording and sentence structure extensively to ensure the new version is distinct from the original. The article is:*

Table 2: Different prompts used for rewriting Urdu articles while maintaining the core narrative. The word "article" was replaced with "headline" in all prompts for *Short* datasets.

best ones using a grid search over various hyperparameters. Among all the models, the Linear SVM achieved the best results.

XLM-R XLM-RoBERTa-base was selected due to its capabilities in multilingual understanding and classification tasks. Considering the dataset size ranging from 1.3k-10k examples and computational resources, we choose to fine-tune XLM-R instead of current large language models like Llama3.1-8B, to avoid overfitting. Manzoor et al. (2024) fine-tuned LLaMA3-8B on a dataset < 10k using both LoRA and full parameter fine-tuning and showed much lower accuracy on empathy score prediction compared with using RoBERTa embedding. We fine-tuned XLM-R using a learning rate of 2×10^{-5}, weight decay of 0.01, and 10 epochs. The parameter *load_best_model_at_end* was set to `True` to retrieve the best model from all epochs.

4.2 Hierarchical Fake News Detection

Baseline results shown in Table 3 demonstrate that performance for the *machine true* and *machine fake* classes is consistently worse than for the *human true* and *human fake* counterparts. This indicates that machine-generated news is ineffectively detected using multiclass classification. This inspires us to separate the four-class fake news classification task into two sequential subtasks: (i) predicting whether a given sentence is written by machine or human (machine-generated text detection), where more machine-generated text detection data from other domains and languages can be leveraged to improve the accuracy, and (ii) determining whether an article or a headline is fake or true (fake news detection), which the current model excels at.

To this end, we proposed a hierarchical method that breaks the multiclass problem into two subtasks: machine-generated text detection and fake news detection. The goal is to break a complex task into two simpler subtasks and improve the performance of each by making use of the data curated for each subtask, ultimately enhancing the overall results, as shown in Figure 2.

We adapted the training labels of four datasets to meet the subtask requirements. For machine-generated text (MGT) detection, we need labels of 'Human' and 'Machine', while 'Fake' and 'True' are for fake news detection. Given the better performance of XLM-R and to ensure a fair comparison with the baseline, we used XLM-R for fine-tuning both two subtasks. The hyperparameters are consistent with those described in training the baseline models. These models are trained and optimized on the validation data accordingly. At inference, both models predict their respective labels for the test data, and these are concatenated and transformed back into the four labels.

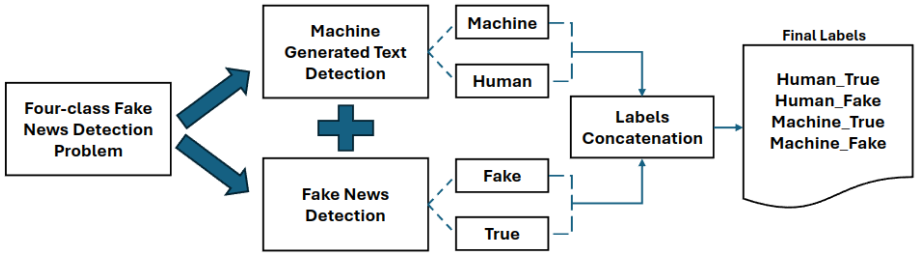


Figure 2: Proposed Hierarchical Fake News Detection Architecture

5 Experiments

This section presents various experimental results and interesting findings from the analysis.

5.1 Four-class vs. Hierarchical Detection

Experimental Setup: Various experiments were conducted to compare the performance of the proposed hierarchical fake news model against the four-class baseline models. These include (i) performance on the four datasets, (ii) comparison of models trained on datasets in the *long* category versus those for the *short* category, and (iii) performance of the model trained for all four datasets combined. A withheld test portion (described in Section 3.1) was used to measure performance for all models. Table 3 summarizes the results obtained, including overall accuracy and F1 score for the four classes. Cross-domain evaluation is performed, which includes how the models trained on one type of dataset perform on the other datasets.

Individual Datasets Results in Table 3 can be analysed to compare the performance of hierarchical fake news detection models trained on individual datasets (1 to 4) and tested on their respective test splits, against those for baseline models. It is clear the new model consistently outperforms the baselines across all four datasets, both for accuracy and the F1 scores for each class. The only exception is dataset 1, where the proposed model ranks a close second to the F1 score of the *human true* class but surpasses it in all other F1 scores and for overall accuracy. Another key improvement is the reduced gap between F1 scores of *human fake* and *machine fake*, as well as *human true* and *machine true*. Although baseline models show a significant difference in F1 scores between *Human* and *Machine* for fake news detection, the proposed hierarchical model largely bridges this gap, achieving almost identical performance, demonstrating the efficacy of the hierarchical classification approach.

Dataset	Model	HF	HT	MF	MT	Acc
Dataset1	*LSVM*	0.73	0.61	0.64	0.52	0.63
	XLM-R-base	0.83	**0.71**	0.77	0.69	0.75
	Hierarchical	**0.85**	0.69	**0.8**	**0.74**	**0.77**
Dataset2	*LSVM*	0.82	0.6	0.77	0.53	0.71
	XLM-R-base	0.93	0.66	0.88	0.7	0.82
	Hierarchical	**0.93**	**0.8**	**0.9**	**0.77**	**0.87**
Dataset3	*LSVM*	0.89	0.87	0.86	0.85	0.87
	XLM-R-base	0.91	0.91	0.88	0.89	0.9
	Hierarchical	**0.96**	**0.95**	**0.92**	**0.91**	**0.94**
Dataset4	*LSVM*	0.56	0.59	0.3	0.42	0.48
	XLM-R-base	0.76	0.73	0.58	0.65	0.68
	Hierarchical	**0.85**	**0.85**	**0.74**	**0.79**	**0.81**
Short	XLM-R-base	0.88	0.68	0.83	0.72	0.78
	Hierarchical	**0.93**	**0.85**	**0.91**	**0.86**	**0.89**
Long	XLM-R-base	0.89	0.88	0.74	0.77	0.82
	Hierarchical	**0.94**	**0.94**	**0.89**	**0.9**	**0.92**
All	XLM-R-base	0.89	0.77	0.83	0.74	0.81
	Hierarchical	**0.91**	**0.85**	**0.88**	**0.83**	**0.87**

Table 3: Accuracy (Acc) and F1-score over four labels on four individual datasets (top four rows) and their combinations: Short=1+2, Long=3+4, and All=1+2+3+4. **HF**: Human Fake, **HT**: Human True, **MF**: Machine Fake, **MT**: Machine True.

Combinations of Datasets This section presents the results for different dataset combinations: *long datasets (dataset 3+4)*, *short datasets (dataset 1+2)*, and *all datasets combined*. Similar training steps and hyperparameters were used, with training splits of datasets 1 and 2 combined for the *short dataset* model, datasets 3 and 4 combined for the *long dataset* model, and all datasets combined for the *all datasets* model. The baseline LSVM, due to consistently lower performance, is excluded from further experiments. The bottom three rows

of Table 3 summarize the performance of models trained on *short*, *long*, and *all* datasets combined. As expected, the proposed model outperforms the baseline across all combined datasets. Secondly, the proposed models narrow the gap between the F1 scores of *Human* and *Machine* in fake news detection compared to the baseline. In comparing performance, models trained on shorter datasets outperform those trained on longer datasets. This is likely because machine detection is easier on longer datasets, as GPT-4o's rephrasing of longer articles results in higher token variation, allowing the classifier to better distinguish between machine and human text. Similarly, for the all-dataset trained model, the proposed model outperforms the baseline by 6% in accuracy and achieves higher F1 scores for machine-generated fake news detection.

Cross-domain Evaluation Cross-domain evaluation was conducted by testing a model trained on one dataset's training set with test splits from other datasets, resulting in a total of 49 evaluations. For brevity, the results presented are only for accuracy. Figures 3a and 3b display the heatmaps of accuracy values for all 49 combinations of proposed and baseline models, respectively. The y-axis represents the training splits and the x-axis represents the test splits of the datasets. The overall trends of both heatmaps are similar, with higher values along the diagonal and lower values in the non-diagonal entries, except for the last entry of the all-dataset-trained model (which is not cross-domain, obviously). This indicates the models do not generalize well for out-of-domain data. However, shorter datasets yield relatively better generalization. This trend is less pronounced for longer datasets. Using our proposed method, the model trained on dataset 3 achieves only 32% accuracy when tested on dataset 4, while the model trained on dataset 4 yields 48% when tested on dataset 3, showing poor performance.

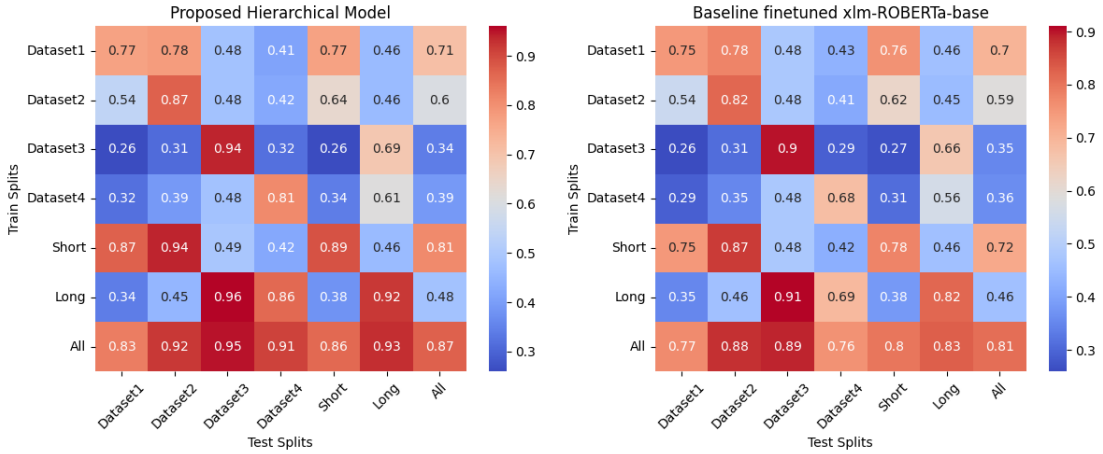
5.2 Analysis

This section offers an analysis of some of the interesting results and includes possible reasons for low accuracy in cross-domain evaluation when short datasets are used for training with long datasets for testing, and vice versa. It also analyses the results of the experiment aimed at enhancing the model's performance on the first dataset, which exhibited the lowest performance among the individual dataset models.

Low Accuracy in Cross-domain Evaluation using *Short* for Training and *Long* for Testing. Figure 4 shows the confusion matrices of the model trained on dataset 1, tested on datasets 3 and 4 respectively. Notably, the matrices exhibit almost no correct predictions for the *machine fake* and *machine true* classes. Interestingly, despite overall incorrect predictions, *machine true* is mostly misclassified as *machine fake*, and *human true* as *human fake*, suggesting that the machine-generated text detection component performs well on both long datasets. The reason the fake news detection module fails, in this case, can be attributed to a simple observation: for short datasets, there is a significant difference in average token count between true and fake classes, with fake articles having more tokens, as shown in Table 1. This may inadvertently cause the model to treat text length as a key feature, resulting in all long articles being classified as fake.

Low Accuracy in Cross-domain Evaluation using *Long* for Training and *Short* for Testing. The models trained on long-category datasets exhibit different behavior when predicting on short datasets. Figure 5 shows the confusion matrix of the two long datasets-trained models given test splits from Dataset 1. Notably, the machine-generated text detection module performs less effectively than in the previous case, with values scattered across the confusion matrix. Secondly, unlike the short datasets, the average tokens for true and fake classes are similar preventing the model from using length as a distinguishing parameter during training. Consequently, the fake news detection module does not classify all short texts as fake. For the model trained on dataset 3, the *human true* class performs poorly, with all samples misclassified as either *machine true* or *machine fake*. In contrast, the model trained on dataset 4 shows a more dispersed confusion matrix with most samples being classified into one of the following three classes: *human true*, *machine true*, or *machine fake*. While precision for *human fake* is high, recall is low, making it less useful. Overall, this model appears to produce somewhat random predictions, likely due to being trained on the smallest dataset among the four.

Improvement in Dataset 1 by Data Augmenting in MGT Detection. Among the individual datasets, dataset 1 has the poorest performance, achieving an accuracy of only 77%. Closer in-



(a) Proposed Model (b) Baseline Model

Figure 3: Cross-domain evaluation results in terms of *Accuracy*

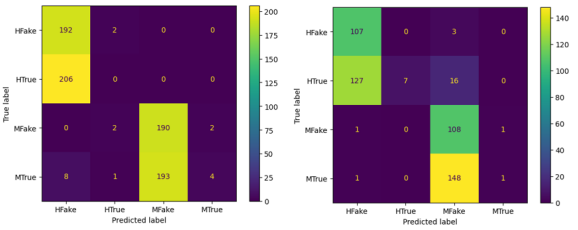


Figure 4: Confusion matrix of testing on long datasets using a model trained on dataset1. **Left:** Test Split Dataset 3 (Long) and **Right:** Test Split Dataset 4 (Long)

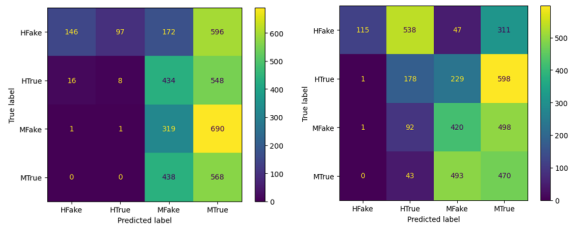


Figure 5: Confusion matrix of testing on dataset 1 using model trained **Left:** Train Split Dataset 3 (Long) and **Right:** Train Split Dataset 4 (Long)

spection reveals that the subpar performance of the machine-generated-text detection module affects overall results. This may be because when GPT-4o rephrases short texts, like those in dataset 1, it makes minimal changes, making it challenging for the model to learn distinguishing features. To test the hypothesis that enhancing machine-generated-text detection would improve overall results, the Urdu subset of a publicly available machine-generated text detection dataset M4 (Wang et al., 2024c) was augmented, and the model was re-trained. This led to a 3% improvement in the accuracy of the MGT module, which subsequently boosted the overall accuracy of the model trained on dataset 1 by 4%. This emphasizes the importance of enhancing machine-generated text detection for the four-label fake news detection, especially for datasets with short texts.

6 Conclusion and Future Work

In this work, we introduced a four-class Urdu fake news detection task and presented the first publicly available datasets for this task. We proposed a hierarchical approach that breaks down the four-class problem into machine-generated-text detection and fake news classification. Experiments demonstrate that our approach consistently enhances the accuracy compared to baseline methods and demonstrates robustness across unseen domains. Moreover, the proposed method effectively bridges the gap between the F1-score of the *machine true* and *human true* classes, as well as *machine fake* and *human fake* classes, thereby improving the identification of machine-generated fake news. in addition, data augmentation for the machine-generated text (MGT) module improved MGT accuracy and thus enhanced the overall performance for four-class fake news detection.

For future work, we will explore methods mitigating the classifier from learning length as a feature during training. Additionally, experiments with other multilingual LLMs could further enhance the performance of fake news detection models. Exploring domain adaptation techniques to improve generalization across diverse datasets and integrating explainability methods to understand model decisions are also interesting.

Limitations

We acknowledge certain limitations in this work that can be addressed in future research. First, the reliance on publicly available datasets may limit the diversity and richness of the training data, potentially affecting the generalizability of our model. This could lead to suboptimal performance when applied to real-world scenarios where misinformation varies widely in style and content. Secondly, the TFIDF features used for the LSVM classifier may not be the most optimal for fake news detection. Alternative features, such as those derived from the News Landscape (NELA), could enhance performance, but their implementation requires considerable effort, particularly for the Urdu language. Third, the model may inadvertently learn to rely on text length as a distinguishing feature, which could skew predictions, especially with varying lengths of articles. This tendency was observed during the analysis of our results, indicating that further refinement is necessary to mitigate this issue. Finally, the machine-generated text detection (MGT) module primarily addresses a subset of machine-generated content, potentially missing other forms of automated misinformation. Future work could focus on expanding the MGT module to encompass a broader range of machine-generated texts.

Ethical Statement and Broad Impact

Ethical Statement We recognize that our approach to fake news detection involves the use of machine-generated text, which may inadvertently incorporate biases present in the training data or models. Given the potential for misinformation to influence public opinion and societal well-being, it is crucial to emphasize the importance of human oversight in the evaluation of our system's outputs. We advocate for the involvement of human reviewers, particularly in sensitive contexts, to ensure responsible decision-making and to mitigate the risk of misclassification.

Broader Impact This work has the potential to significantly enhance the field of fake news detection, particularly for low-resource languages like Urdu. By providing publicly available datasets and a robust hierarchical approach, this research will empower journalists, researchers, and the general public to identify and combat misinformation more effectively. The proposed methodology can be adapted for various applications, including integration into news platforms and social media, thereby facilitating the identification of misleading information and contributing to the overall integrity of public discourse. Ultimately, this work aims to foster a more informed society by improving the tools available for discerning fact from fiction in the rapidly evolving digital landscape.

References

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. Detection of online fake news using n-gram analysis and machine learning techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1*, pages 127–138. Springer.

Muhammad Pervez Akhter, Jiangbin Zheng, Farkhanda Afzal, Hui Lin, Saleem Riaz, and Atif Mehmood. 2021. Supervised ensemble learning methods towards automatically filtering urdu fake news within social media. *PeerJ Computer Science*, 7:e425.

Maaz Amjad, Grigori Sidorov, and Alisa Zhila. 2020a. Data augmentation using machine translation for fake news detection in the urdu language. In *Proceedings of the 12th language resources and evaluation conference*, pages 2537–2542.

Maaz Amjad, Grigori Sidorov, Alisa Zhila, Helena Gómez-Adorno, Ilia Voronkov, and Alexander Gelbukh. 2020b. "bend the truth": Benchmark dataset for fake news detection in urdu language and its evaluation. *Journal of Intelligent & Fuzzy Systems*, 39(2):2457–2469.

Manoj Kumar Balwant. 2019. Bidirectional lstm based on pos tags and cnn architecture for fake news detection. In *2019 10th International conference on computing, communication and networking technologies (ICCCNT)*, pages 1–6. IEEE.

Alberto Barrón-Cedeno, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56:1849–1864.

Bhavika Bhutani, Neha Rastogi, Priyanshu Sehgal, and Archana Purwar. 2019. Fake news detection using sentiment analysis. In *2019 twelfth international conference on contemporary computing (IC3)*, pages 1–5. IEEE.

Muhammad Shoaib Farooq, Ansar Naseem, Furqan Rustam, and Imran Ashraf. 2023. Fake news detection in urdu language using machine learning. *PeerJ Computer Science*, 9:e1353.

Sheetal Harris, Jinshuo Liu, Hassan Jalil Hadi, and Yue Cao. 2023. Ax-to-grind urdu: Benchmark dataset

for urdu fake news detection. In *2023 IEEE 22nd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 2440–2447. IEEE.

Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2022. Faking fake news for real fake news detection: Propaganda-loaded training data generation. *arXiv preprint arXiv:2203.05386*.

Soufia Kausar, Bilal Tahir, and Muhammad Amir Mehmood. 2020. Prosoul: a framework to identify propaganda from online urdu content. *IEEE access*, 8:186039–186054.

Stephan Lewandowsky, Ullrich KH Ecker, and John Cook. 2017. Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of applied research in memory and cognition*, 6(4):353–369.

Muhammad Arslan Manzoor, Yuxia Wang, Minghan Wang, and Preslav Nakov. 2024. Can machines resonate with humans? evaluating the emotional and empathic comprehension of LMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14683–14701, Miami, Florida, USA. Association for Computational Linguistics.

Ahmadreza Mosallanezhad, Kai Shu, and Huan Liu. 2021. Generating topic-preserving synthetic news. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 490–499. IEEE.

OpenAI. 2024. Hello gpt-4o.

Michael Orlov and Marina Litvak. 2019. Using behavior and text analysis to detect propagandists and misinformers on twitter. In *Information Management and Big Data: 5th International Conference, SIMBig 2018, Lima, Peru, September 3–5, 2018, Proceedings 5*, pages 67–74. Springer.

Marialaura Previti, Victor Rodriguez-Fernandez, David Camacho, Vincenza Carchiolo, and Michele Malgeri. 2020. Fake news detection using time series and user features classification. In *Applications of Evolutionary Computation: 23rd European Conference, EvoApplications 2020, Held as Part of EvoStar 2020, Seville, Spain, April 15–17, 2020, Proceedings 23*, pages 339–353. Springer.

Shaina Raza and Chen Ding. 2022. Fake news detection based on news content and social contexts: a transformer-based approach. *International Journal of Data Science and Analytics*, 13(4):335–362.

Somya Ranjan Sahoo and Brij B Gupta. 2021. Multiple features based approach for automatic fake news detection on social networks using deep learning. *Applied Soft Computing*, 100:106983.

Kai Shu, H Russell Bernard, and Huan Liu. 2019. Studying fake news via network analysis: detection and mitigation. *Emerging research challenges and opportunities in computational social network analysis and mining*, pages 43–65.

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

Jinyan Su, Claire Cardie, and Preslav Nakov. 2023a. Adapting fake news detection to the era of large language models. *arXiv preprint arXiv:2311.04917*.

Jinyan Su, Terry Yue Zhuo, Jonibek Mansurov, Di Wang, and Preslav Nakov. 2023b. Fake news detectors are biased against texts generated by large language models. *arXiv preprint arXiv:2309.08674*.

Edson C Tandoc Jr, Zheng Wei Lim, and Richard Ling. 2018. Defining "fake news" a typology of scholarly definitions. *Digital journalism*, 6(2):137–153.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. 2024a. SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Mexico City, Mexico. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024b. M4GT-bench: Evaluation benchmark for black-box machine-generated text detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3964–3992, Bangkok, Thailand. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024c. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407, St. Julian's, Malta. Association for Computational Linguistics.

Xiaoxin Yin, Jiawei Han, and Philip S Yu. 2007. Truth discovery with multiple conflicting information providers on the web. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1048–1052.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.