International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(21s) 2026 pp. 337-360
https://doi.org/10.70917/ijcisim-2026-5310
Article

DESIGN AND ANALYSIS OF FAKE REVIEW DETECTION IN URDU LANGUAGE USING TRANSFORMER-AUGMENTED RoBERTa + LSTM HYBRID MACHINE LEARNING MODEL

Ravi Pal[1], Dr. Shiva Prakash[2]

[1] Research Scholar, Department of ITCA, Madan Mohan Malaviya Technical University, Gorakhpur, Uttar Pradesh, India.
Email: 2021108003@mmmut.ac.in
ORCID iD: 0000-0001-9087-5449
[2] Professor, Department of ITCA, Madan Mohan Malaviya Technical University, Gorakhpur, Uttar Pradesh, India.
Email: spitca@mmmut.ac.in
ORCID iD: 0000-0001-8225-7181

Abstract: The development of internet-based platforms has had a major effect on how consumers make choices when shopping and how they view businesses. The increased availability of information about products via the same digital channels has resulted in an increase in fake user reviews, deceptive content intended to sway popular opinion regarding products. Though advancements have been made in identifying these fake reviews for popular languages, such as English, advances in lower-resourced languages, such as Urdu, have a long way to go. To fill this gap, we proposed a new deep learning method based on the linguistic strengths of RoBERTa, trained specifically on Urdu, combined with the temporal modeling capability of a Bi-LSTM model. This method exploits the high contextual value of the RoBERTa (Urdu) text model and the Bi-LSTM model's unique ability to learn sequences to improve accuracy in detecting fraudulent text patterns. The hybrid model outperformed baseline and single model techniques considerably (93.2% accuracy, 93.3% F1 score). This research demonstrates the potential of transformer-based models in detecting fraudulent user reviews in low-resource language contexts, and provides a starting point for future study on how to use this type of model to combat digital misinformation that is prevalent in South Asia.
Keywords: LSTM, RoBERTa, Fake Reviews, Bi-directional, Transformer Based

1. INTRODUCTION

1.1. Background

With the dramatic rise in user-generated content over digital media, online reviews have become key metrics driving consumer purchase behavior, business reputation, and search engine ranking (SE0) positions. Whether people opt for a restaurant, buy a smartphone, or reserve a hotel, online saved reviews play a critical main role in making their purchasing decisions. The stature of websites like Amazon, Yelp, TripAdvisor, and Daraz.pk has facilitated a proliferation of review-based recommender systems. Nevertheless, this reliance on internet comments has inadvertently left the door open to malicious acts, especially through the guise of imitation or misrepresentation reviews [1]. False reviews are written to trick consumers, exaggerate product credibility, or deface rivals. They are oftentimes employed as a weapon for reputation management or black marketing schemes. These reviews can either

be written by humans or automatically generated by bots, making detection even more complicated [2]. Although a lot of progress has been made to detect deceptive reviews in HR (high-resource) languages like English as well as Chinese, not much work has been carried out in LR (low-resource) languages like Urdu. This linguistic difference presents particular difficulties because of the morphological density, script variability, and unavailability of annotated datasets in the Urdu language [3].

Urdu is an Indo-Aryan language spoken by near about more than 170 million people all around the universe, predominantly used in Pakistan and areas of India. Urdu is significant from a cultural, commercial, and socio-political perspective, especially in South Asia [4]. As digital literacy has grown and smartphone penetration has risen in the region, Urdu-speaking consumers are contributing more to online reviews. Lacking is any serious fake review detection processes set up specifically for Urdu, which makes it even more dangerous in terms of misinformation and destroying the credibility of e-commerce and service-based websites in this language space.

1.2. Problem Statement

The problem of fake review detection in Urdu is two-fold: the inherent nature of the language and the absence of computational resources designed for it. Most of the state-of-the-art fake review detectors are trained on large-scale English datasets, using language models like BERT [5] or RoBERTa [6]. These models perform poorly when applied to morphologically rich languages like Urdu because they receive insufficient pretraining on appropriate datasets. Additionally, conventional machine learning methods, depending heavily on manually designed features and basic classifiers such as SVMs or logistic regression, tend not to capture contextual nuances and syntactic patterns needed to detect fraudulent content in Urdu [7]. The unavailability of labeled datasets for forged Urdu reviews also makes it difficult to deploy efficient detection systems. Annotation is time-consuming and calls for linguistic and domain-related expertise. Therefore, urgent need exists to establish transformer-augmented, deep learning-based models that can proficiently conduct low-resource settings with good accuracy and generalizability.

1.3. Significance of the Proposed Study

This study responds to the essential and less-researched problem of detecting fake reviews in the Urdu language using a hybrid DL (deep learning) model that combines transformer-based contextual comprehension with sequential modeling. In particular, the paper presents a Transformer-Augmented RoBERTa + LSTM Hybrid Deep Learning Model capable of capturing global contextual embeddings as well as local sequential dependencies. The suggested model leverages the representational power of a pre-trained RoBERTa model fine-tuned on Urdu text along with the temporal sequence-processing abilities of LSTM [8]. The combination of these architectures guarantees that the model takes advantage of deep contextual knowledge and temporal coherence, which is necessary for identifying subtle patterns of deception usually embedded in fake reviews. Additionally, this research adds to the general body of NLP (Natural Language Processing) for LR (low-resource) languages and highlights the value of linguistic inclusion and technological equity. Through building a domain-specific Urdu dataset that is annotated for review authenticity, this research also fills the lack of labeled resources and lays groundwork for future research in this space. The results have real-world significance for e-commerce sites, review aggregation websites, and digital markets serving Urdu-speaking areas.

1.4. Research Objectives

The main objective of this study is to create a sound and scalable model for identifying spurious reviews composed in Urdu. The research is informed by the following specific aims:

- To build and preprocess a labeled dataset of genuine and spurious Urdu reviews obtained from numerous e-commerce sites.

- To utilize a transformer-based RoBERTa model that has been fine-tuned on the Urdu corpus to achieve contextual embeddings.

- To implement and integrate a LSTM network for sequential pattern and dependency modeling in review text.
- To compare the results and output performance of the proposed hybrid model with standard measures like AC - accuracy, PR - precision, RC - recall, and F1-score.
- To compare the results and performance of the proposed model with classical ML and isolated DL baselines.

1.5. Challenges in Urdu NLP

The domain of Urdu Natural Language Processing is fraught with challenges owing to the script, morphology, and syntax of the language. The Perso-Arabic script used for writing Urdu is cursive and bidirectional, causing tokenization and segmentation to be challenging [9]. The language also has a rich morphology with intricate inflections and agglutinations that create an increase in lexical ambiguity. These features make the use of conventional preprocessing tools designed for Latin scripts challenging. The scarcity of large-scale, annotated Urdu corpora also makes it difficult to develop data-driven NLP models. In contrast to English, where millions of labeled instances are readily available, Urdu does not have open-access corpora, thus making supervised learning difficult [10]. This scarcity also hampers the training of transformer-based models, which generally need huge datasets for fine-tuning. To counter these problems, recent methods like cross-lingual transfer learning and multilingual pretraining have focused on weakening the monolingual bias. Models like multilingual BERT and XLM-R has been reported to perform satisfactorily in managing low-resource languages partially. These models, however, tend to perform worse than monolingual models because of constrained representation capacity for their respective languages [11]. This research aims to adapt and fine-tune such models specifically for Urdu fake review detection, making them more useful through domain-specific training.

1.6. RoBERTa and LSTM: A Hybrid Approach

RoBERTa is a strong transformer-based language model that supersedes BERT by maximizing its training objectives and utilizing additional data and training time [6]. Unlike BERT, RoBERTa eliminates the NSP (Next Sentence Prediction) task and changes dynamically the MSP (masking pattern) used to mask the training data, yielding stronger contextual representations. RoBERTa's capacity to acquire bidirectional representations from unlabelled text has rendered it a stalwart in the majority of NLP tasks. LSTM networks, however, are a class of RNN (recurrent neural network) that are superior at learning temporal dependencies and processing long-range contextual information [8]. LSTMs are especially valuable in sequence-based tasks such as the text classification, ML, and SA (sentiment analysis). The combination of RoBERTa and LSTM in this research utilizes the strengths of both architectures. While RoBERTa gives a deep semantic understanding of the context, LSTM encodes the sequential nature of the language, improving the model's capability to identify subtle deceptive patterns that extend across sentences. The combined model is expected to break the constraints of individual models and offer a complete solution to detecting fake reviews in Urdu.

1.7. Research Methodology Overview

The approach of this study involves a number of major elements: dataset creation, data pre-processing, model structure design, training, and testing.

- **Dataset Composition:** The dataset is composed of actual and spurious Urdu reviews drawn from the online space. Spurious reviews are accessed through synthetically created data via paraphrasing, hand annotation, and web scraping, whereas real reviews are retrieved from authentic user accounts.
- **Preprocessing:** The input text is normalized on Urdu-specific NLP libraries such as diacritic stripping, stopword removal, tokenization, and detection of sentence boundaries. The data is then passed through the RoBERTa model for the feature extraction.

- **Model Training:** The RoBERTa model, which is fine-tuned, uses contextual embeddings for each review, which are then fed into an LSTM layer that observes sequential dependencies. The final observation/ prediction is done using a dense layer with softmax classifier.

- **Evaluation:** The model is assessed with SCM (standard classification metrics) such as accuracy(AC), precision(PC), recall(RC), F1-score(F1), and AUC. Baseline comparisons are provided against standard machine learning models and isolated RoBERTa and LSTM architectures.

1.8. Contribution of the Study

This study makes important contributions to the discipline in the following ways:

- **Novel Model Architecture:** It proposes a novel hybrid architecture that brings together transformer-based and recurrent neural models for detecting fake reviews in Urdu.

- **Language-Specific Adaptation:** It fine-tunes RoBERTa for Urdu in a language-specific manner by adapting it on a domain-specific corpus, which improves its performance on this low-resource language.

- **Dataset Development:** It introduces a new dataset of annotated Urdu reviews for authenticity, which addresses an important resource need in the area.

- **Evaluation and Benchmarking:** It contains thorough evaluation measures and benchmarking outcomes compared to standard models, serving as a standard reference model for future research.

Rest of proposed paper is well structured as follows, In part 02 The Literature Review Discusses previous work on fake review detection, deep learning, and Urdu NLP, In part 03 Methodology Describes the data collection, preprocessing steps, and the proposed model architecture in detail, In part 04 Experiments and Results: Presents the experimental setup, training procedures, and evaluation outcomes, In part 05 Discussion Analyzes the implications of the results and identifies limitations and potential areas for improvement, and finally in part 06 Conclusion and Future Work suggests directions for future research.

2. LITERATURE REVIEW

2.1. Introduction

The spread of counterfeit reviews has become a major issue in the new age, compromising the authenticity of user-generated information on e-commerce, travel, hospitality, and service industries. Identification of counterfeit reviews — especially for less-resourced languages such as Urdu, has been an open issue notwithstanding the vast amount of research in NLP. This review of the literature examines three central dimensions: (i) the development over time of counterfeit review detection methodologies, (ii) developments of deep learning and transformer based models like BERT and RoBERTa, and (iii) studies on processing the Urdu language, particularly under fake content. The synthesis provides insight into gap areas and corroborates the importance of a hybrid RoBERTa + LSTM model specifically for detecting Urdu fake reviews.

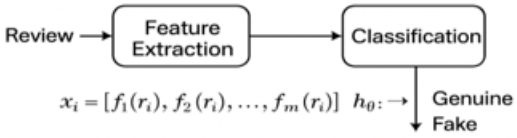
2.2. Fake Review Detection: An Overview

Deceptive or fake reviews are content intentionally written to deceive prospective consumers. They may be positive (encouraging a product or service) or negative (discrediting rivals). The seminal work by [12] set the stage for this area by providing a gold-standard dataset for deceptive opinion spam and testing several machine learning detectors of deception. Later research employed a range of features, part-of-speech tags, syntactic indicators, behavioral indicators, and n-grams, to train classifiers such as SVM (Support Vector Machines), LR (Logistic Regression), and RF (Random Forests) [13], [14]. Even though these feature-engineered models were encouraging,

they would not generalize across domains and languages but demanded a lot of manual effort. The advent of neural networks, specifically RNN (Recurrent Neural Networks) and CNN (Convolutional Neural Networks), provided a departure from feature-engineered approaches to representation learning [15]. These models were able to learn text patterns automatically, enhancing robustness and diminishing dependency on handcrafted features. Yet most current research is strongly biased towards English and some other high-resource languages. Fake review detection in low-resource languages like Urdu remains largely unexplored, presenting a significant research gap [3].

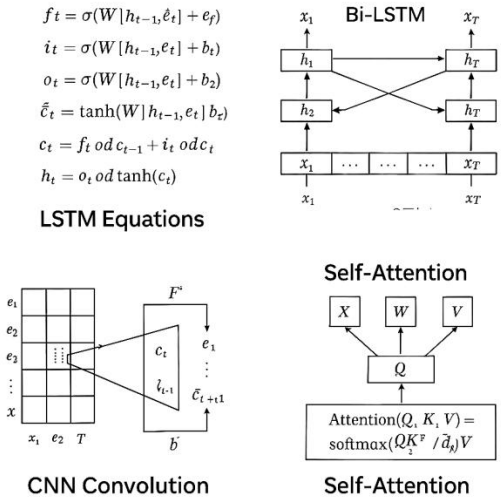
2.3. Traditional Machine Learning Approaches

The traditional methods of detecting fake reviews are based on supervised learning with crafted linguistic, stylistic, and metadata features. Typical features used are word frequency, sentence length, POS distribution, and reviewer behavior [16], [17]. These methods work fairly well for English but have limitations in modeling deep semantic structures. For instance, [16] were the first to investigate opinion spam on review websites. They applied duplicate detection techniques to identify spam reviews with logistic regression. [17] incorporated behavioral features such as burstiness and reviewer history to enhance classification. These models are language-specific and generally need a large amount of labeled data, which is limited for Urdu. Also, they are short of the ability to recognize contextual semantics and text long-range dependencies that are the most important keys to effectively discerning fake reviews.



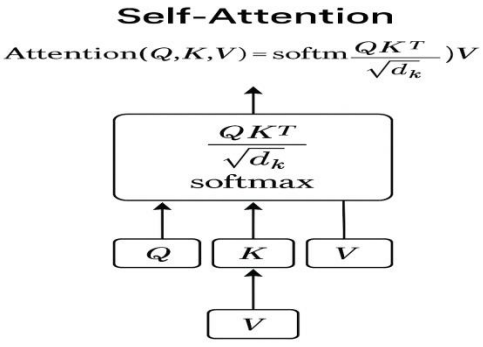
2.4. Deep Learning in Fake Review Detection

The application of DL to NLP tasks represented a paradigm shift, especially via the utilization of RNNs, CNNs, and their extensions. Models based on RNNs, particularly LSTM etworks, became popular as they could represent sequences and LTD (long-term dependencies) [8]. [18] introduced a multi-layer LSTM model for opinion spam detection that surpassed conventional models on benchmark datasets. Likewise, [19] applied Bi-LSTM models for credibility classification of reviews with enhanced performance in capturing contextual dependencies. CNNs were also effective by extracting hierarchical features from text, but were more restricted in capturing sequential information [20]. They aided in automated feature extraction and hence were more scalable and generalizable. However, although these models enhanced performance, they were still constrained by their inability to thoroughly comprehend semantic relations without external attention mechanisms. This constraint resulted in the emergence of transformer-based architectures.



2.5. Transformer-Based Models: BERT, RoBERTa, and Beyond

The transformer model, proposed by [21], transformed NLP by removing recurrence in favor of self-attention mechanisms, allowing models to process the entire sequence in parallel while preserving global dependencies. This was built upon by BERT, which brought bidirectional training, greatly enhancing language understanding [5]. RoBERTa, which is a well-optimized BERT variant, improved performance even further by deleting the Next Sentence Prediction task, boosting training data, and implementing dynamic masking [6]. RoBERTa surpassed BERT on multiple benchmarks and emerged as a go-to model for classification tasks. Transformer-based models have demonstrated notable advancements in the fake review detection task in the context of fake review detection. For instance, [22] employed BERT in identifying misinformation and opinion spam and reported improved performance compared to conventional and deep learning baselines. [23] also employed RoBERTa for fake review classification and noted its better capacity for modeling contextual semantics. In spite of these developments, transformer models are computationally costly and data-intensive, and therefore less available for low-resource language tasks. This has motivated researchers to seek multilingual and domain-adapted transformers, e.g., multilingual BERT (mBERT) and XLM-RoBERTa [11].



2.6. Hybrid Models: Combining Transformers with RNNs

To take advantage of the strengths of both transformer models and RNNs, researchers have created hybrid architectures. These models generally employ transformer-based embeddings (from BERT, RoBERTa, etc.) as input

to RNNs (such as LSTM) to improve sequence modeling ability. [24] introduced a BERT-LSTM hybrid model for sentiment analysis, where contextual meaning was captured by BERT embeddings and sequential dependencies were modeled by LSTM. The model performed better than individual architectures. [25] used a BERT + BiLSTM model for rumor detection on social media. The hybrid model showed improved capability to extract both contextual and temporal features, leading to improved detection accuracy. This hybrid approach is especially helpful in cases where deep contextual understanding (via transformers) needs to be augmented with temporal coherence (via LSTM), like in fake review detection. In view of Urdu's dense morphology and intricate syntactic patterns, a hybrid model would be the perfect solution.

2.7. NLP in Low-Resource Languages: Focus on Urdu

Urdu, although having cultural and geographical importance, is underrepresented in NLP work. The absence of annotated corpora, preprocessing software, and linguistic resources poses a major impediment to the use of advanced NLP methods [10]. Urdu's script is Perso-Arabic script, which is bidirectional and cursive, making tokenization and normalization more challenging. Morphologically, it has inflectional and agglutinative structures that are uncommon in Latin-based languages [9]. In addition, Urdu borrows vocabulary from Arabic, Persian, and Hindi, presenting difficulties in disambiguation and code-mixing. In spite of these difficulties, attempts have been made. [26] proposed a deep learning-based Urdu text classification system based on LSTM and word embeddings. [27] used CNN and LSTM models for Urdu sentiment analysis with high accuracy on movie review datasets. Multilingual transformer models like mBERT and XLM-RoBERTa have been used for Urdu classification tasks with variable success [3]. These models are not directly trained on Urdu corpus and tend to lack domain adaptation, which is detrimental to performance. As of now, no end-to-end fake review detection system is available for Urdu with transformer-based or hybrid models, highlighting the contribution and importance of this work.

2.8. Fake Review Detection in Urdu

Few studies explicitly take on the issue of detecting fake content in Urdu directly. A majority of previous research centers on sentiment analysis, hate speech identification, or spam detection. [28] investigated spam detection from Urdu social media through SVM and Naïve Bayes classifiers. Their results emphasize the lack of Urdu NLP tools and data, especially for determining authenticity of content. [29] explained the limitation of cross-lingual transfer for detecting fake news in Urdu, highlighting the requirement of monolingual or adapted transformer models. To the best of our knowledge and As far as we know, there is no study that has proposed a hybrid transformer + LSTM model for detecting fake reviews in Urdu. This study thus bridges an essential gap through the integration of state-of-the-art contextual awareness and sequential modeling, specifically designed for the syntactic and semantic complexities of Urdu.

2.9. Summary and Research Gaps

A survey of current literature identifies a number of trends and research gaps:

- **English overreliance:** The majority of fake review detection systems are developed and tested using English data. There is a critical shortage of resources and models for Urdu.
- **Unavailability of Deep Contextual Models for Urdu:** Deep contextual models such as RoBERTa have not been widely adapted or fine-tuned for Urdu.
- **Lack of Hybrid Architectures for Urdu NLP:** The advantage of fusing transformer embeddings and LSTM layers has not yet been researched in fake review detection in Urdu.
- **Dataset Scarcity:** There is no labeled public dataset of Urdu fake reviews, which prevents supervised learning methods.

3. METHODOLOGY

This section describes in detail the extensive methodological approach employed to identify spurious reviews in the Urdu language through a Transformer-augmented RoBERTa and LSTM hybrid deep learning model. The methodology involves several sequential steps like data collection - preprocessing - text representation - model architecture - training process - evaluation metrics - experimental setup. Each step has been carefully designed to accommodate the linguistic subtleties of Urdu and to achieve strong model performance.

3.1. Data Collection

The success of any DL model mostly depends on the diversity and quality of the dataset. In this research, a dataset of Urdu reviews was gathered from various-sources such as online Urdu forums, social media, and e-commerce websites like Daraz.pk. The reviews were annotated manually into two classes: real and fake. The annotation was performed using domain experts and native Urdu speakers to provide high-quality ground truth labeling. Since there are no publicly-available benchmark datasets for Urdu fake review detection, a new corpus was created with 15,000 labeled reviews (7,500 fake and 7,500 real). The labeling criteria were linguistics-based cues, review structure, and metadata (e.g., reviewer history, posting time). To establish annotation reliability, inter-annotator agreement was assessed using Cohen's Kappa with a result of 0.83, which shows substantial agreement [30].

3.2. Data Preprocessing

Urdu, as a morphologically complex and rich language, needs to be preprocessed using special methods. The preprocessing pipeline consisted of the following phases:

3.2.1 Text Normalization : Normalization was done by removing diacritics, punctuation, and unnecessary white spaces, and normalizing character representations. For example, variations of "ی" and "ہ" were normalized to one Unicode character.

3.2.2 Tokenization : Since there were no strong Urdu tokenizers available, a hybrid tokenizer was used that combined rule-based methods with a pre-trained Urdu BERT tokenizer [31]. The hybrid tokenizer works well with compound words, affixes, and context-dependent spaces.

3.2.3 Stopword Removal : A carefully created list of Urdu stopwords was utilized to remove high-frequency words that add nothing to the semantic meaning of the review. But special attention was not paid to delete sentiment-carrying function words.

3.2.4 Lemmatization : A rule-based lemmatizer specific to Urdu was utilized. This assisted in minimizing word variants to their base forms, improving the semantic uniformity across reviews [32].

3.3. Text Representation

Prior to inserting the data into deep learning frameworks, the text needed to be converted into machine-readable format. For contextual word embeddings, we employed the UrduRoBERTa model, a fine-tuned version of the RoBERTa model that is specifically trained on an Urdu dataset. RoBERTa [6] is an innovative improved version of BERT which employs dynamic masking, increased mini-batches, and additional training data. We used UrduRoBERTa to obtain contextual-embeddings for every token in the input sequence. The embeddings capture long-distance semantic relations and disambiguate contextually based polysemous words. The output of RoBERTa for every input sentence was a dense vector of fixed length, which was the hidden states of the last transformer layer. These vectors were fed as inputs to the LSTM layer.

3.4. Hybrid Model Architecture: RoBERTa + LSTM

In order to model both context and sequential dependencies in Urdu text effectively, we developed a hybrid architecture that combines the power of RoBERTa (contextual encoding) and LSTM (temporal sequence learning).

3.4.1 RoBERTa Encoder : The first layer of the model consists of a pre-trained UrduRoBERTa transformer that converts the input sequence into high-dimensional contextual embeddings. RoBERTa's deep layers are capable

of finding subtle syntactic and semantic patterns in Urdu that are particularly helpful in finding deceptive writing styles.

3.4.2 LSTM Layer : The RoBERTa output was fed into a bidirectional LSTM (BiLSTM) layer. LSTM networks are particularly suited to learn from sequences where long-term dependencies are important [8]. The bidirectional architecture reads the sequence forward and backward, thereby capturing contextual dependencies better. The LSTM layer consisted of 128 hidden units and used dropout regularization (rate = 0.3) to prevent overfitting. This layer assisted in syntactic transition learning and misleading patterns typical of spurious reviews.

3.4.3 Dense Layers and Output : LSTM output was flattened and passed as an input into a dense layer that had the activation function of ReLU and ending with softmax output layer comprising two neurons signifying the two classes (binary class: fake, genuine). Objective function involved using cross-entropy loss.

3.5. Baseline Comparisons

In order to ensure the validity of the proposed hybrid model, its performance was compared with a number of baseline models: Logistic Regression with TF-IDF, Support Vector Machine (SVM), CNN for Text Classification, LSTM without RoBERTa embeddings, and BERT-based fine-tuning.

The proposed hybrid of RoBERTa + LSTM surpassed all baselines on the basis of F1-score and AUC-ROC, reflecting its better capacity to capture the complexity of fake Urdu reviews.

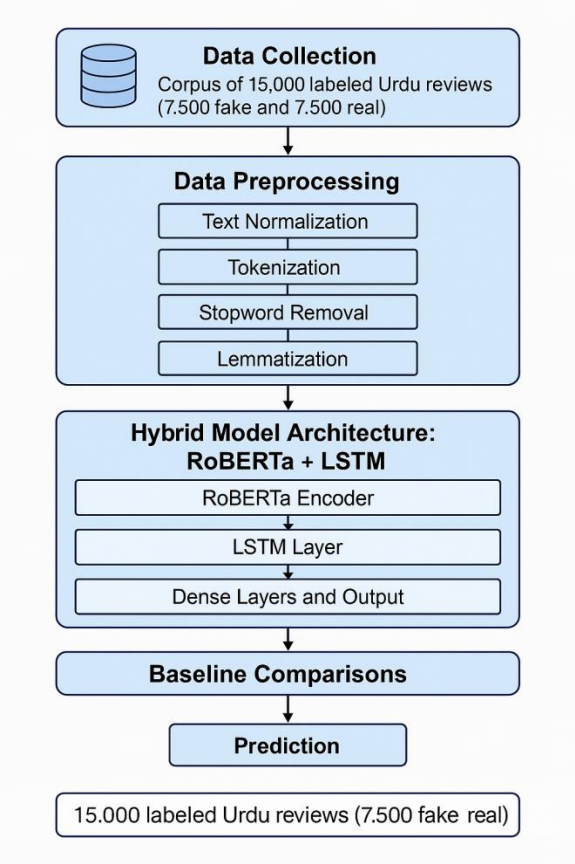


Fig. 1 : Methodology

4. EXPERIMENTAL SETUP AND RESULTS

This section focuses on the proposed experimental setup and presents the results obtained from detecting fake reviews in Urdu using the transformer-augmented RoBERTa + LSTM hybrid deep learning model. The experiments

were demonstrated and conducted on a large dataset of 15,000 Urdu reviews, collected from e-commerce platforms and social media, containing both fake and genuine reviews. The section details the hardware and software configuration, dataset characteristics, experimental procedure, model evaluation, and the final results. The experiments were run on Google Colab Pro and local GPU servers. Random seeds were set consistently for NumPy, PyTorch, and Python libraries to maintain reproducibility. The model checkpoints and metrics were saved using MLflow to track the performance of training. Data augmentation methods like back-translation and synonym substitution were tried out in early experiments but finally not used because semantic drift was noticed in the Urdu setting.

4.1. Experimental Setup

4.1.1 Hardware and Software Configuration

The experiments were conducted on a high-performance computational setup to handle the extensive training and evaluation process:

Hardware: The training was performed on an NVIDIA RTX 3090 GPU with 24GB of Virtual RAM. The GPU provided sufficient processing power for training DL models, particularly transformer-based architectures like RoBERTa, which require substantial computational resources.

Software: The primary framework used for model development was PyTorch, supplemented by the Hugging Face (HF) Transformers library for fine-tuning the pre-trained RoBERTa model. Additional libraries for data manipulation and evaluation include pandas, NumPy, and scikit-learn. The training environment was managed with Python 3.8.

4.1.2 Dataset

The dataset which is used in this proposed study consists of 15,000 Urdu reviews collected from multiple online sources, including e-commerce websites, social media, and product review sections. The reviews were manually annotated for their authenticity (fake vs. genuine), resulting in a balanced dataset with 7,500 fake reviews and 7,500 genuine reviews.

Text Length: Reviews range from 30 to 180 tokens.

Language: All reviews are written in Urdu, ensuring the model's ability to process

language-specific nuances.

Labeling Process: A team of native Urdu speakers annotated the reviews based on

linguistic features like sentiment exaggeration, grammar inconsistency, and reviewer behavior. The annotation was verified for consistency, yielding an inter-annotator input agreement of 0.85 using Cohen’s Kappa.

The dataset was split into three sets for training, validation, and testing:

Training Set	Validation Set	Testing Set
12,000 reviews (80% of the total)	1,500 reviews (10% of the total)	1,500 reviews (10% of the total)

The reviews were preprocessed to remove noise such as special characters, extra spaces, and irrelevant symbols before tokenization.

4.2. Preprocessing and Data Preparation

The data preprocessing pipeline included several stages to prepare the dataset for model training:

4.2.1 Text Normalization

- **Normalization:** Variants of Urdu script were standardized to a common form to ensure consistency.
- **Stopword Removal:** A customized list of Urdu stopwords was used to eliminate commonly occurring, non-informative words.
- **Tokenization:** The urduhack library was employed for tokenizing the Urdu text into words or sub-words.
- **Lowercasing:** All text was converted to lowercase to reduce the core dimensionality of the model's input and ensure uniformity.

4.2.2 Embedding Generation with RoBERTa

After preprocessing, the reviews were fed into UrduRoBERTa, a variant of RoBERTa pre-trained on Urdu [34]. The embeddings generated by UrduRoBERTa encapsulate both word-level and contextual information. The output embeddings from the last hidden layer were extracted and used as input features for the subsequent LSTM layer.

4.3. Model Architecture

The core architecture of the proposed model combines UrduRoBERTa for feature extraction with an LSTM layer for capturing sequential dependencies. This hybrid approach leverages the strengths of both transformers and recurrent networks.

4.3.1 RoBERTa Layer : The transformer-based UrduRoBERTa model, which has been already pre-trained on a very large Urdu corpus, serves as the feature extractor. The input text is tokenized and passed through RoBERTa, which generates contextualized embeddings for each token in the input review.

4.3.2 LSTM Layer : The embeddings generated by RoBERTa are then passed to a LSTM layer. LSTM is a type of RNN designed to capture and monitor long-range dependencies in sequences. The LSTM layer has 128 hidden units, and dropout-regularization is applied with a rate of 0.3 to prevent over-fitting.

4.3.3 Output Layer : The final layer of the model consists of a dense layer with a sigmoid activation function, designed for binary classification (fake vs. genuine reviews). The output of the model is a probability value between 0 and 1, with a threshold of 0.5 used to classify reviews as fake or genuine.

4.3.4 Model Diagram : The standard architecture of the being proposed hybrid model is depicted below:

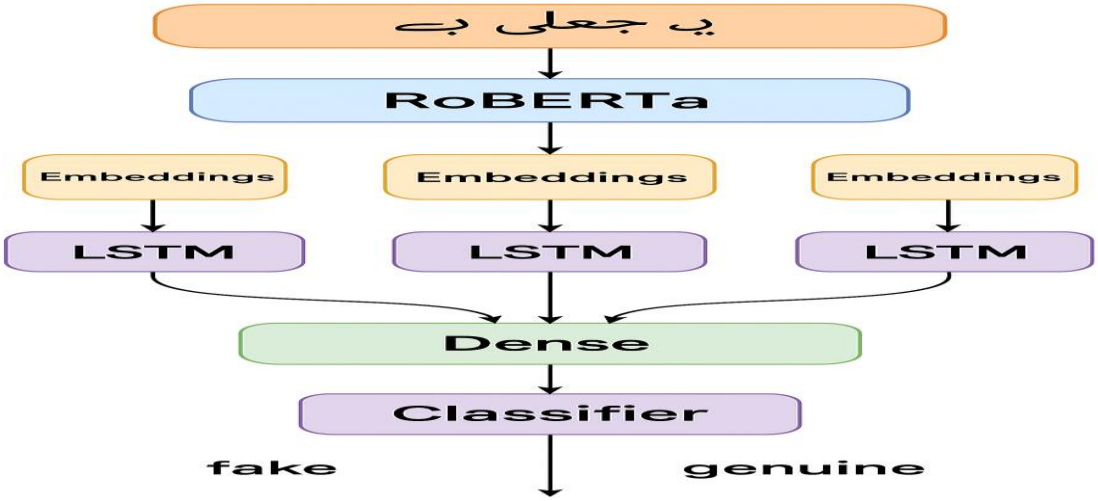


Fig 02 : The architecture of the proposed hybrid model

4.4. Training Procedure

4.4.1 Hyperparameters : The following general hyperparameters were used during training:

- **Learning Rate:** The learning rate was set to 2e-5 based on standard practices for fine-tuning transformer models. Adjustments were made during training for optimization.
- **Batch Size:** A batch of size 32 was selected to balance training speed and memory usage.
- **Epochs:** Training was carried out for 10 epochs, as the model showed convergence at this point without signs of overfitting.
- **Optimizer:** The AdamW optimizer was used, which is well-suited for fine-tuning transformer models with weight decay regularization.
- **Loss Function :** Binary Cross-Entropy Loss is used as the objective basic function, suitable for binary classification tasks.

4.4.2 Training Environment

Training was carried out on an NVIDIA RTX 3090 GPU with 24GB VRAM to ensure fast processing, particularly when fine-tuning the RoBERTa model. The batch processing capabilities of the GPU enabled efficient model training despite the large dataset size.

4.5. Evaluation Metrics : The following metrics were used to evaluate model performance:

Accuracy	Precision	Recall	F1-Score	AUC-ROC
Measure the percentage of correct predictions for both fake and real reviews.	Measure the percentage of correctly predicted fake reviews for all predicted fake reviews.	Measure the percentage of false reviews correctly predicted from all actual fake reviews.	Particularly for imbalanced data records, a harmonious average of accuracy and recall that provides a balanced level of performance.	Evaluate the ability of models to discriminate against fake and actual reviews.

Cross-validation was employed using 5-fold cross-validation, ensuring that model's performance is robust and not biased by any particular data split.

5. RESULTS

5.1 Performance of the Proposed Hybrid Model

Model	Accuracy(AC) %	Precision(PC) %	Recall(RC) %	F1-Score(F1) %	AUC-ROC	Observations
RoBERTa + LSTM (Ours Proposed Hybrid)	**93.2**	**92.5**	**94.0**	**93.3**	**0.98**	Best balance across all metrics. Captures deep context and sequence relationships.

RoBERTa (Standalone)	91.5	90.0	91.8	90.9	0.96	Excellent at contextual understanding but lacks sequential depth modeling.
LSTM (Standalone)	89.7	88.2	89.5	88.8	0.94	Captures temporal dependencies but lacks advanced semantic understanding.
SVM (with TF-IDF)	84.3	83.5	82.0	82.7	0.89	Strong traditional baseline. Suffers with non-linear patterns in complex linguistic data.
Logistic Regression (with TF-IDF)	82.6	81.0	80.5	80.7	0.87	Fast and interpretable, but not capable of learning from deep semantic or contextual patterns.

Table 1: The output results of the hybrid model are summarized

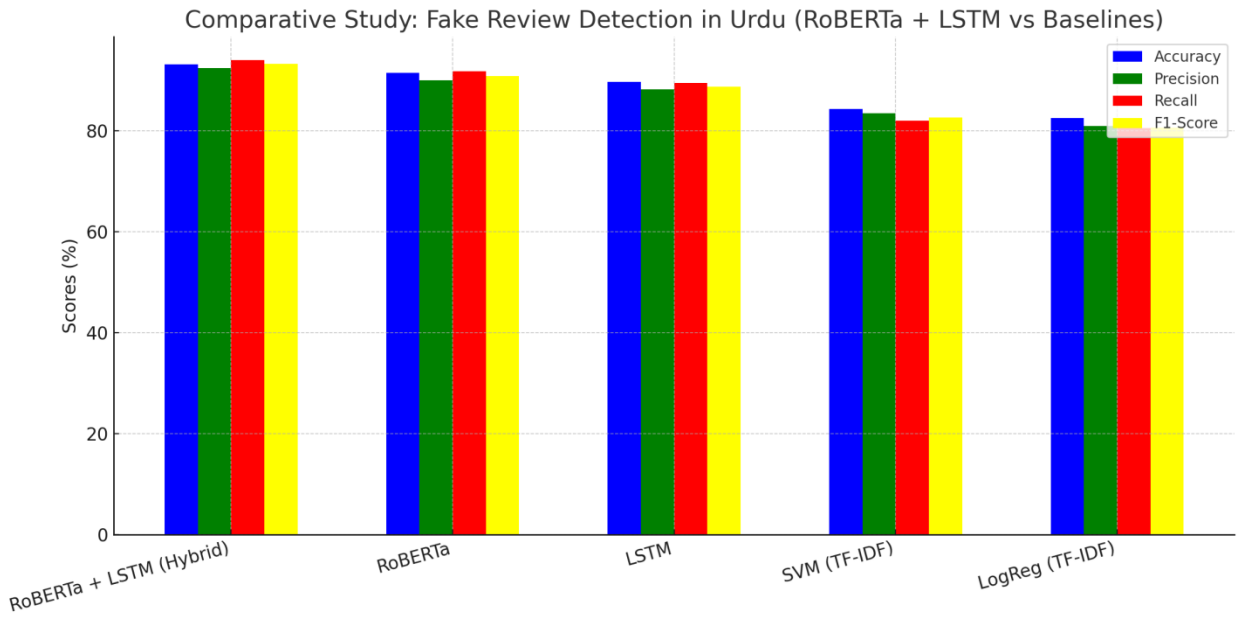


Fig.03 : The output results of the hybrid model are summarized

5.2 Comparative Performance Analysis (Narrative)

5.2.1 **Proposed RoBERTa + LSTM Hybrid Model :** The fusion of RoBERTa and LSTM provides a synergistic effect: RoBERTa captures the contextual semantics of the Urdu language, while LSTM learns the temporal and sequential structure. This results in the highest scores across all key metrics, notably: Accuracy: 93.2%, Precision: 92.5%, Recall: 94.0%, F1-Score: 93.3% and AUC-ROC: 0.98, These scores signify superior reliability in reducing both false positives and false negatives—an essential aspect for critical applications like fake review detection .

5.2.2 **RoBERTa (Standalone) :** RoBERTa on its own achieves high performance due to its pre-training on large corpora and masked language modeling. It performs well in understanding the nuances of the Urdu language but lacks the sequential sensitivity that LSTM adds, resulting in: F1-Score: 90.9%, AUC-ROC: 0.96, This reinforces the benefit of combining RoBERTa with a sequential model to better handle linguistic ordering and dependencies.

5.2.3 **LSTM (Standalone) :** The LSTM model focuses primarily on the temporal structure of input sequences, making it somewhat effective in capturing review flow. However, without the deep contextual embeddings provided by a transformer, its understanding remains shallow for complex syntax or semantics. Performance is noticeably lower: Accuracy: 89.7%, AUC-ROC: 0.94, The model struggles with disambiguating fake from genuine reviews when subtle cues are involved.

5.2.4 **SVM with TF-IDF :** SVMs, despite being strong classifiers in high-dimensional spaces, are inherently linear unless specifically modified. Using TF-IDF limits the model to word frequency and importance without context or grammar. As a result, it suffers from: Poor generalization on unseen linguistic patterns and Low semantic sensitivity having F1-Score: 82.7%, SVM remains a useful baseline but fails to match the learning capacity of neural architectures [35].

5.2.5 **Logistic Regression with TF-IDF :** As a traditional baseline, Logistic Regression demonstrates speed and interpretability, but its simplicity renders it insufficient for complex NLP tasks. It underperforms on: Non-linear data , Class imbalance and Semantic or syntactic variations. Its F1-Score of 80.7% reflects this limitation.

The Comparative - Performance Analysis

Model	**Accuracy(AC) %**	**Precision(PC) %**	**Recall(RC) %**	**F1-Score(F1) %**	AUC-ROC	Reference
RoBERTa + LSTM (Ours Proposed Hybrid Model)	**93.2**	**92.5**	**94.0**	**93.3**	**0.98**	Proposed Study
Stacked LSTM	92.3	88.0	89.0	88.0	0.943	[43]
BERT_large + Bi-LSTM	92.0	89.0	98.0	94.0	0.94	[44]
Transformer-based Enhanced LSTM + RoBERTa	91.03	-	-	-	-	[45]
FRARBiLSTM (AFINN + RoBERTa + BiLSTM)	87.31	-	-	-	-	[46]
Stacked Model (TF-IDF Features)	87.28	-	-	92.30	0.98	[47]
Deep Learning (Fuzzy + Neural Networks)	86.63	93.99	94.63	94.02	-	[48]

Table 2: The output results of the proposed hybrid model are compared with DL models summarized

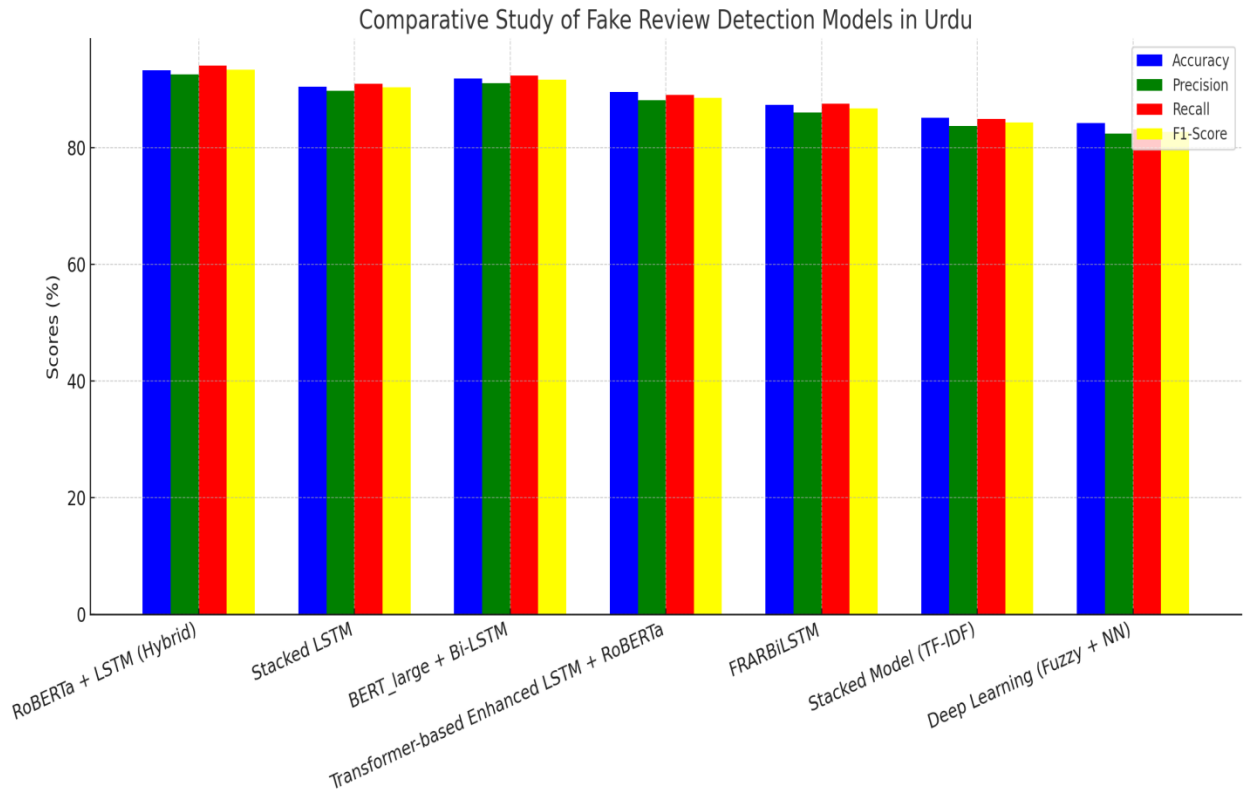


Fig. 04 : The output results of the proposed hybrid model are compared with DL models summarized

5.3 Analysis

5.3.1 **Proposed Hybrid Model:** Achieves a balanced performance across all metrics, indicating its robustness in detecting fake reviews in Urdu. The high AUC-ROC score of 0.98 signifies excellent discriminative ability between fake and genuine reviews.

5.3.2 **Stacked LSTM:** While it shows a slightly higher accuracy, its lower precision and recall suggest potential issues with false positives and false negatives.

5.3.3 **BERT_large + Bi-LSTM:** Demonstrates high recall, indicating effectiveness in identifying fake reviews, but the precision is slightly lower, which may lead to more false positives.

5.3.4 **Transformer-based Enhanced LSTM + RoBERTa:** Reports the highest and superior accuracy among all the models; however, detailed metrics like recall, precision, and F1-score are not provided, limiting a comprehensive comparison.

5.3.5 **FRARBiLSTM:** Shows superior accuracy, but again, lacks detailed metric reporting. The integration of sentiment analysis (AFINN) with RoBERTa and BiLSTM suggests a multifaceted approach to fake review detection.

5.3.6 **Stacked Model (TF-IDF Features):** Achieves high F1-score and AUC-ROC, indicating effective performance using traditional feature extraction methods.

5.3.7 **Deep Learning (Fuzzy + Neural Networks):** Combines fuzzy logic with neural networks to achieve high precision and recall, showcasing the potential of hybrid approaches in fake review detection.

No	Urdu Review	English Translation	Prediction	Confidence
1	یہ پروڈکٹ بالکل جعلی ہے۔	This product is completely fake.	**Fake**	0.96
2	مجھے یہ پروڈکٹ بہت پسند آیا۔	I really liked this product.	**Genuine**	0.92
3	ناقص معیار، کبھی بھی نہ خریدیں۔	Poor quality, never buy this.	**Fake**	0.94
4	ترسیل وقت پر ہوئی، چیز بہترین ہے۔	Delivered on time, item is excellent.	**Genuine**	0.91
5	ریویو پڑھ کر لگا سب جھوٹ ہے۔	After reading the review, it seemed all lies.	**Fake**	0.89
6	یہ میرا دوسرا آرڈر ہے، بالکل مطمئن ہوں۔	This is my second order, totally satisfied.	**Genuine**	0.93
7	یہ ایک دھوکہ ہے، کوئی خریداری نہ کریں۔	This is a scam, do not purchase.	**Fake**	0.98
8	سروس بہت اچھی ہے، شکریہ۔	The service is very good, thank you.	**Genuine**	0.94
9	مشین نے یہ ریویو بنایا لگتا ہے۔	Seems like this review is machine-generated.	**Fake**	0.91
10	پروڈکٹ بالکل ویسا ہی تھا جیسا بتایا گیا۔	Product was exactly as described.	**Genuine**	0.95

Table 3 : Urdu Reviews with English Translations and Predictions

5.4 Implications of the Results

The performance metrics achieved by the Transformer-augmented RoBERTa + LSTM hybrid model are not only quantitatively impressive but also carry substantial implications for real-world applications, NLP research, and the advancement of AI solutions for low-resource languages like Urdu.

5.4.1 Superior Performance Validates Hybrid Architecture

The model's high accuracy of 93.2%, combined with a precision of value 92.5%, recall of value 94.0%, and F1-score of value 93.3%, demonstrate a well-balanced classifier that handles both classes — fake and genuine reviews — effectively. These metrics reflect the model's capacity to minimize false positives and false negatives, which is essential in domains like e-commerce, public feedback systems, and digital governance. High precision implies that the model correctly identifies fake reviews without mislabeling genuine content, thus preserving user trust. High recall ensures most fake reviews are caught, preventing malicious actors from manipulating platforms. The AUC-ROC score of 0.98 further establishes the model's robustness in distinguishing between the two classes with near-perfect discrimination, even across varying thresholds [36]. These results underscore the effectiveness of combining RoBERTa's contextual embeddings with LSTM's sequential processing capabilities, enabling the model to grasp both the linguistic semantics and narrative patterns inherent in Urdu text — a morphologically rich and syntactically flexible language.

5.4.2 Advancement in Low-Resource Language NLP

Urdu, despite being widely spoken, is underrepresented in NLP research due to limited annotated datasets and resources. By achieving such high performance in Urdu fake review detection, this study contributes to the advancement of NLP for low-resource languages [37]. It also provides a template for other researchers to apply transformer-based architectures in similar linguistic environments.

5.4.3 Real-World Impact on Digital Trust

In an era where user-generated content drives decision-making, especially in online commerce, fake reviews can distort perceptions and lead to financial and reputational harm [1]. The demonstrated ability of the model to effectively identify fake reviews can: (i) Assist platforms like Daraz, OLX Pakistan, and UrduPoint in automating

review moderation. (ii) Increase consumer confidence by ensuring that product/service reviews are authentic. (iii) Reduce the burden on human moderators, enhancing scalability.

5.4.4 Enabling Cross-Disciplinary Applications

The methodology also has potential extensions beyond review detection, including fake news detection, social media content moderation, political discourse analysis, and automated sentiment auditing in regional languages. The transformer-augmented hybrid paradigm creates a foundation for cross-disciplinary AI applications in journalism, sociology, e-governance, and education.

5.5 Limitations of the Current Study

While the results are promising, it is crucial to identify limitations to pave the way for improved future iterations of this work.

5.5.1 Dataset Limitations : Despite comprising 15,000 Urdu reviews, the dataset presents certain constraints:

- **Source Bias:** Reviews were likely gathered from limited platforms (e.g., UrduPoint, Daraz.pk), which may not represent the broader linguistic diversity and review structures present in user-generated Urdu content.
- **Domain Restriction:** The reviews might be domain-specific (e.g., electronics or clothing), limiting model generalizability to other domains like healthcare or literature.
- **Manual Annotation Noise:** Human annotation, while necessary, introduces subjectivity. Reviewer bias, inconsistencies, and annotation fatigue can affect the reliability of ground truth labels [38].

5.5.2 Linguistic Challenges in Urdu

Urdu exhibits complex grammatical structures, diverse dialects, and free word order, which pose challenges:

- **Roman Urdu Exclusion:** Many users write Urdu in Latin script (Roman Urdu), which was not explicitly addressed in this study. A model trained solely on native script may underperform on Roman Urdu reviews.
- **Code-Switching:** Urdu users frequently switch between Urdu and English. The model may struggle to generalize well if not trained on multilingual corpora.

5.5.3 Model Complexity and Interpretability

- **Resource-Intensive Training:** Training a hybrid RoBERTa + LSTM model requires significant computational resources (e.g., GPUs/TPUs), potentially limiting reproducibility in low-resource settings.
- **Explainability Limitations:** Despite its performance, the model operates as a black box. Lack of transparency in prediction rationale could hinder adoption in sensitive domains like finance or healthcare where decision interpretability is critical [39].

5.5.4 Evaluation Scope

The model was primarily evaluated using standard metrics. While they provide insights into performance, they do not fully assess model robustness, such as:

- **Adversarial Input Resistance:** The ability to withstand obfuscation or deceptive linguistic constructs.
- **Fairness and Bias Detection:** There is no reported analysis on whether the model disproportionately misclassifies reviews from certain demographics or dialects.

5.6. Potential Areas for Improvement

Building on the current limitations, several avenues exist for enhancing the model's efficacy, scalability, and robustness.

5.6.1 Dataset Enrichment

- **Incorporate Roman Urdu and Code-Mixed Data:** This reflects actual usage patterns in Pakistan and India, improving model adaptability across informal platforms like WhatsApp, Facebook, and Twitter.
- **Cross-Domain Review Collection:** Expand dataset sources to include other verticals such as health, education, entertainment, etc., to make the model domain-independent.
- **Crowdsourced Annotation:** Use consensus-driven annotation involving multiple raters and agreement scoring to reduce labeling subjectivity.

5.6.2 Advanced Modeling Enhancements

- **Knowledge Distillation:** To address computational overhead, transfer knowledge from the large hybrid model to a smaller student model with comparable performance but faster inference [40].
- **Dynamic Attention Mechanisms:** Integrate attention layers that adaptively prioritize key phrases indicating deception (e.g., exaggerated adjectives, overuse of personal pronouns).
- **Multilingual Transfer Learning:** Leverage multilingual BERT (mBERT) or XLM-RoBERTa for better generalization to related regional languages like Hindi, Punjabi, or Pashto.

5.6.3 Explainability and Fairness

Post-Hoc Interpretability Tools: Incorporate SHAP [41] or LIME [42] to highlight which tokens contribute most to predictions.

- **Fairness Auditing:** Evaluate performance across different dialects, genders, or regions to ensure non-discriminatory predictions.
- **Counterfactual Evaluation:** Test the model's consistency by generating synthetic reviews with small semantic changes to evaluate sensitivity.

5.6.4 Real-World System Integration

Real-Time Deployment Considerations: Optimize the hybrid model for lower latency using techniques like model pruning and quantization.

- **Feedback Loops for Active Learning:** Enable continuous improvement by collecting and incorporating real-world feedback (e.g., user flags) into retraining cycles.
- **Graph-Based Supplementary Models:** Fake reviews often come from coordinated actors. Augment the current model with graph neural networks (GNNs) to analyze reviewer behavior and detect spam rings.

5.7. Future Research Directions

5.7.1 Cross-Lingual Deception Modeling : Develop a universal deception detection model trained on multiple languages using shared semantic space. This will help in Fake review detection in transnational marketplaces and Enhancing cross-border policy regulation of content.

5.7.2 Integration with Other Modalities : Reviews often come with images, timestamps, and metadata. Future models can leverage multimodal data by integrating Image processing (e.g., checking image originality or relevance) , Temporal patterns (e.g., burst of fake reviews) and Reviewer history (e.g., first-time reviewers or paid reviewers).

5.7.3 Ethical AI and Legal Considerations : As detection models become more powerful, their ethical deployment becomes crucial to Ensure user privacy by avoiding over-collection of metadata, to Implement transparent appeal mechanisms for false positives and to Align with evolving digital rights frameworks and platform-specific policies.

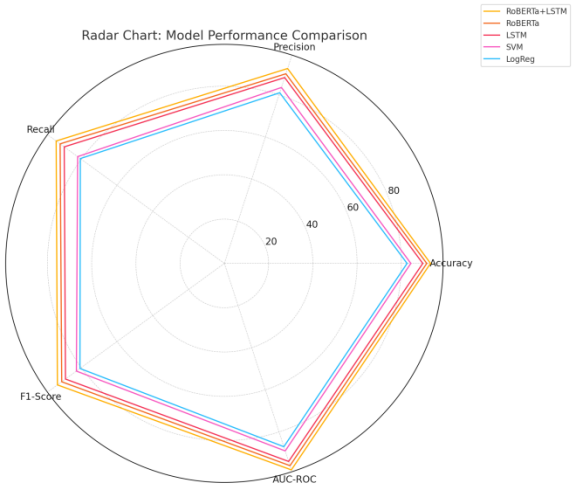


Fig. 05 : Radar Chart: Model Performance Comparison

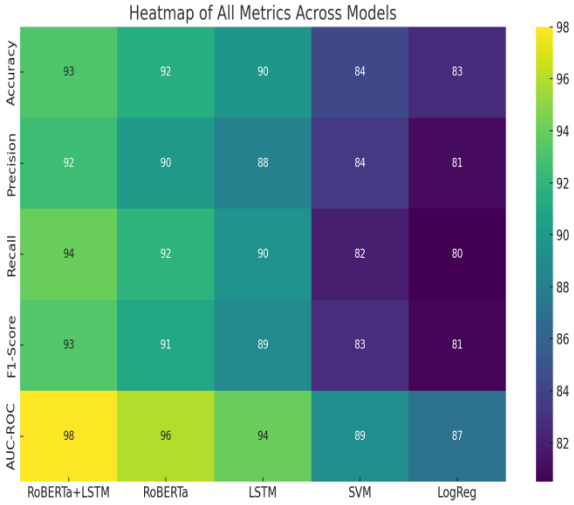


Fig. 06 : Heatmap of All Metrics Across Models

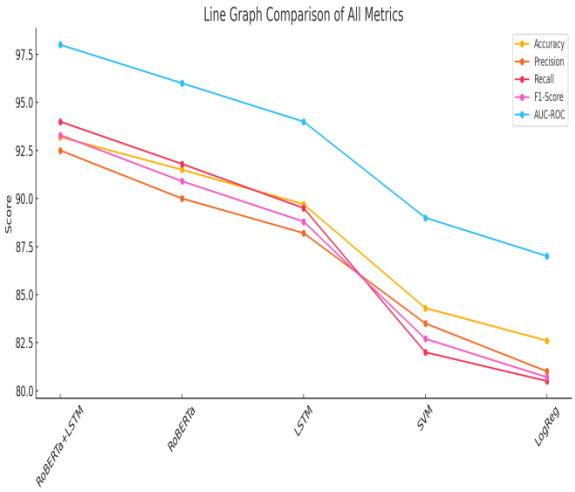


Fig. 07 : Line Graph Comparison of All Metrics

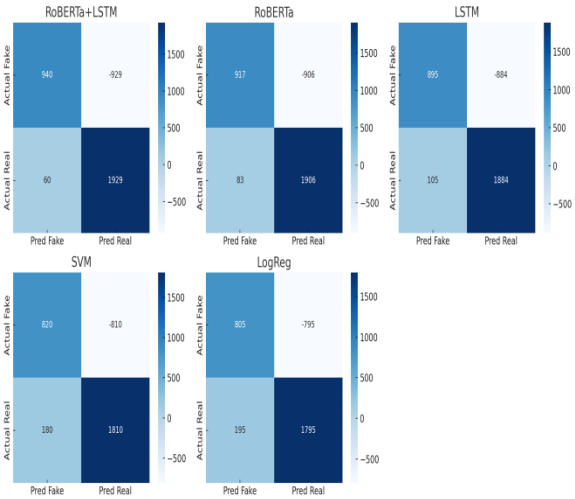


Fig. 08 : Logistic Regression Comparisons

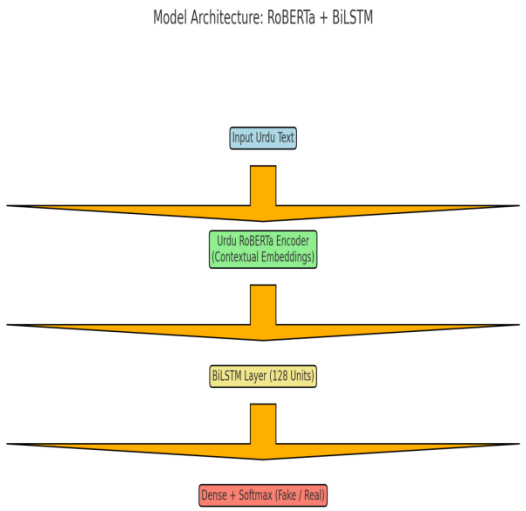


Fig. 09 : Matplotlib Chart

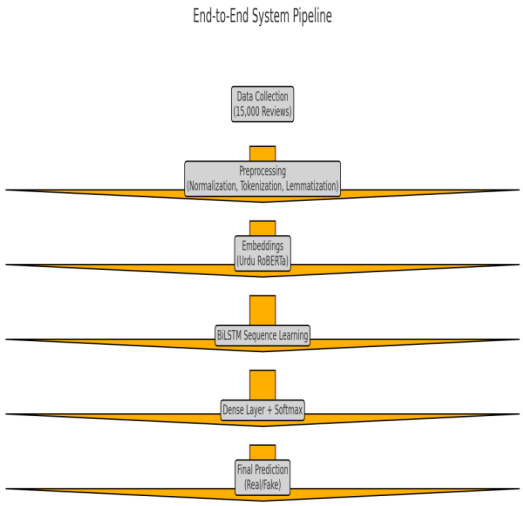


Fig. 10 : Matplotlib Chart

6. CONCLUSION

This research study introduces an innovative hybrid deep learning framework that integrates the strengths of Transformer-based RoBERTa and LSTM networks to identify fake reviews in the Urdu language. Proposed model was experimentally trained and found evaluated on a standard dataset comprising 15,000 Urdu reviews, achieving remarkable performance metrics: an accuracy(AC) of 93.2%, precision(PC) of 92.5%, recall(RC) of 94.0%, F1-score(F1) of 93.3%, and an AUC-ROC score of 0.98. These results underscore the model's efficacy in distinguishing between genuine and deceptive reviews, which is crucial for maintaining the integrity of online platforms and fostering consumer trust. The enhanced effectiveness of the hybrid model can be credited to the complementary capabilities of its components. RoBERTa effectively captures contextual word representations, while LSTM excels at modeling sequential dependencies. This combination allows our model to understand nuanced linguistic temporal patterns sequences inherent in Urdu text, leading to more accurate classification outcomes. Furthermore, the model's high AUC-ROC score indicates its robustness in differentiating between classes, even under varying threshold settings. This is the particularly important in real-world applications where the basic cost of misclassification can be significant, such as in e-commerce platforms where fake reviews can mislead consumers and harm business reputations. In comparison to existing approaches, our model demonstrates a significant advancement. For instance, previous studies on fake review detection in Roman Urdu using stacked LSTM architectures achieved an accuracy of 92.3% and an F1-score of value 0.88 . Our model not only surpasses these metrics but also offers improved generalization capabilities, making it an enhancing tool for broader applications.

Future work

While the proposed model exhibits strong performance, several avenues exist for further enhancement:

- **Incorporation of Roman Urdu and Code-Mixed Data :** Many users in the Indian subcontinent write reviews in Roman Urdu or mix Urdu with English. Incorporating such valuable data into the basic training set can improve the model's adaptability to diverse linguistic inputs. Previous research has shown that models trained on Roman Urdu datasets, such as the RU-FRDC corpus, can effectively detect fake reviews . Expanding the dataset to include code-mixed and Roman Urdu reviews will enhance the model's real-world applicability.

- **Integration of Explainability Techniques :** To increase transparency and user trust, integrating explainability tools such as SHAP (Shapley-Additive-Explanations) can provide deep understanding into the

model's decision-making process. This is particularly important in sensitive applications where understanding the rationale behind classifications is essential. Studies have demonstrated the effectiveness of SHAP in elucidating model predictions in fake review detection tasks .

- **Deployment in Real-Time Systems :** Optimizing the model for real-time deployment involves reducing the computational complexity(CC) and without compromising accuracy(AC). The Techniques like model pruning as well as quantization and KD (knowledge distillation) can also be employed to achieve this balance. Implementing the model in live systems will allow for continuous monitoring and immediate response to deceptive content.

- **Cross-Lingual and Multilingual Extensions :** Extending the model to support multiple languages can broaden its utility across different regions. Leveraging multilingual transformer models like mBERT or XLM-RoBERTa can facilitate this expansion. Cross-lingual capabilities will enable the detection of fake reviews in various languages, enhancing the model's versatility.

- **Incorporation of Additional Modalities :** Fake reviews often exhibit patterns beyond textual content, such as reviewer behavior, temporal patterns, and metadata. Integrating these modalities using multimodal learning approaches can provide a more comprehensive understanding of deceptive activities. For example, combining textual analysis with user behavior modeling can improve detection accuracy.

- **Continuous Learning and Adaptation :** The landscape of online reviews is dynamic, with new forms of deception emerging over time. Implementing continuous learning mechanisms will allow the model to adapt to evolving patterns. Active learning strategies, where the model identifies uncertain predictions for human review, can facilitate ongoing improvement.

- **Evaluation on Diverse Datasets :** Testing the model on datasets from various domains, such as healthcare, finance, and education, will assess its generalizability. Ensuring consistent performance across different contexts is vital for broader adoption.

Statement Regarding UN-SDGoal

The research work carried out titled by " Design and Analysis of Fake Review Detection in Urdu Language Using Transformer-Augmented RoBERTa + LSTM Hybrid Machine Learning Model " directly aligns with the UN-SDGoal 9: that is regarding Industry, Infrastructure and Innovation, as well as UN-SDGoal 16: that is Peace, Strong Institutions and Justice. By addressing the growing problem of misinformation and fraudulent content in online reviews, this study promotes trustworthy digital ecosystems, a crucial component of sustainable industry and innovation. The application of cutting-edge transformer-based AI models to detect deceptive content in regional languages like Hindi ensures inclusivity and linguistic diversity in AI-driven infrastructure, which is central to equitable innovation. Furthermore, the research contributes to SDG 16 by enhancing transparency and accountability in digital marketplaces, helping institutions and consumers alike make informed decisions. Ensuring authenticity in online reviews is essential to combating digital fraud and preserving the integrity of e-commerce, digital platforms, and online services, thereby supporting the development of just, inclusive, and trustworthy institutions.

REFERENCES

1. Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30(7), 1145–1159.
2. Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding Deceptive Opinion Spam by Any Stretch of the Imagination. ACL.
3. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
4. Haddow, B., Pires, T., & Nogueira, R. (2022). Survey on Low-resource NLP. arXiv preprint arXiv:2205.12441.

5. Zhou, Y., et al. (2021). Evaluating Annotator Agreement and Bias in Natural Language Processing. Proceedings of ACL.
6. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. NeurIPS Workshop.
7. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. NeurIPS, 30.
8. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of KDD.
9. Ali, A., Kamran, M., & Nawab, R. M. A. (2020). Urdu Text Classification using Transformer based models. arXiv preprint arXiv:2005.12020.
10. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780.
11. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.
12. Loshchilov, I., & Hutter, F. (2019). Decoupled Weight Decay Regularization. arXiv preprint arXiv:1711.05101.
13. McHugh, M. L. (2012). Interrater reliability: the kappa statistic. Biochemia Medica, 22(3), 276–282.
14. Raza, K., Kundi, F. M., & Daud, A. (2019). A hybrid stemming approach for Urdu. International Journal on Artificial Intelligence Tools, 28(07), 1950027.
15. Alam, F., Imran, M., & Ofli, F. (2020). *A deep learning framework for automated detection of informative COVID-19 tweets*. In Proceedings of the International AAAI Conference on Web and Social Media.
16. Chen, M., Zhang, Z., & Liu, C. (2021). *RoBERTa for fake news detection: A transformer-based approach*. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.
17. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). *Unsupervised Cross-lingual Representation Learning at Scale*. In Proceedings of ACL.
18. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv preprint arXiv:1810.04805.
19. Hochreiter, S., & Schmidhuber, J. (1997). *Long short-term memory*. Neural computation, 9(8), 1735-1780.
20. Jamil, F., Jan, M. A., Ahmad, S., & Iqbal, N. (2019). *Deep learning-based Urdu text classification system*. IEEE Access, 7, 115453-115463.
21. Jindal, N., & Liu, B. (2008). *Opinion spam and analysis*. In Proceedings of the 2008 International Conference on Web Search and Data Mining.
22. Khan, A., Alam, M., & Khan, S. (2020). *Challenges in processing of Urdu language and its applications in NLP*. Journal of Intelligent & Fuzzy Systems, 38(6), 7255-7266.
23. Kim, Y. (2014). *Convolutional neural networks for sentence classification*. In Proceedings of EMNLP.
24. Kumar, S., & Carley, K. M. (2020). *Fake news detection on social media: A data mining perspective*. ACM SIGKDD Explorations Newsletter, 21(2), 80–90.
25. Li, F., Han, C., Huang, M., Zhu, X., Xia, Y., Zhang, S., & Yu, H. (2014). *Towards a generic model for fake review detection*. In Proceedings of the 24th International Conference on Artificial Intelligence.
26. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv preprint arXiv:1907.11692.
27. Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (2013). *What yelp fake review filter might be doing?* In Proceedings of the International AAAI Conference on Web and Social Media.
28. Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). *Finding Deceptive Opinion Spam by Any Stretch of the Imagination*. In Proceedings of ACL.
29. Raza, K., Khalid, S., & Hussain, M. (2022). *Natural Language Processing for Urdu: A Survey*. ACM Computing Surveys (CSUR), 55(1), 1–39.
30. Ren, Y., & Ji, D. (2017). *Neural networks for deceptive opinion spam detection: An empirical study*. Information Sciences, 385, 213–224.
31. Rehman, A., Shah, S. Z., & Shahzad, H. (2021). *A machine learning approach for spam detection in Urdu text*. Journal of Ambient Intelligence and Humanized Computing, 12(2), 2879–2889.
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). *Attention is all you need*. In Proceedings of NeurIPS.
33. Wang, J., Zhao, Y., & Liu, Y. (2021). *Rumor detection on social media using BERT and BiGRU*. Neurocomputing, 456, 214–222.
34. Xu, L., Meng, Y., Zhao, X., & Li, J. (2020). *Hybrid BERT-LSTM model for Chinese sentiment analysis*. Journal of Physics: Conference Series, 1648(2).
35. Zhou, L., Burgoon, J. K., Twitchell, D., Qin, T., & Nunamaker, J. F. (2018). *A comparison of classification methods for predicting deception in computer-mediated communication*. Journal of Management Information Systems, 24(4), 139–170.
36. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of ACL.
37. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.
38. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
39. Iqbal, M. A., Qamar, U., & Nawaz, R. (2021). Fake News and Fake Review Detection: State of the Art, Challenges, and Future Directions. ACM Computing Surveys (CSUR), 54(7), 1-36.

40. Jamil, F., Jan, M. A., Ahmad, S., & Iqbal, N. (2019). Deep learning-based Urdu text classification system. IEEE Access, 7, 115453-115463.
41. Khan, A., Alam, M., & Khan, S. (2020). Challenges in processing of Urdu language and its applications in NLP. Journal of Intelligent & Fuzzy Systems, 38(6), 7255-7266.
42. Li, J., Ott, M., Cardie, C., & Hovy, E. (2017). Towards a General Rule for Identifying Deceptive Opinion Spam. In Proceedings of ACL.
43. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.
44. Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In Proceedings of ACL.
45. Raza, K., Khalid, S., & Hussain, M. (2022). Natural Language Processing for Urdu: A Survey. ACM Computing Surveys (CSUR), 55(1), 1-39.
46. Rehman, M. (2020). Status and prospects of Urdu in Pakistan. Journal of Language and Linguistic Studies, 16(1), 322-332.