# Hate Speech Detection in Urdu News: A Web-Scraped Dataset and ML/DL Classification Framework

Atif Saeed
Department of Data Science
FAST–NUCES, Islamabad
atif.saeed@isb.nu.edu.pk

Talha Aslam
Department of Data Science
FAST-NUCES, Islamabad
i248067@isb.nu.edu.pk

Muhammad Ahmer
Department of Data Science
FAST–NUCES, Islamabad
i247602@isb.nu.edu.pk

***Abstract*—This paper presents a complete framework for hate speech detection in Urdu news using a web-scraped dataset and machine-learning (ML) and deep-learning (DL) models. Articles were collected from four major Urdu news outlets and processed through an Urdu-focused preprocessing and lexicon-based labeling pipeline. Classical ML models using TF–IDF features achieved the best results, with XGBoost and the Stacking Classifier reaching an accuracy and F1-score of 0.90. Deep-learning models, including BiLSTM, CNN, a Hybrid CNN–BiLSTM, and BiLSTM with Attention, showed competitive recall but lower overall accuracy due to dataset size constraints. The findings demonstrate that ensemble ML approaches outperform DL models for structured news text, while attention-based architectures offer potential for improved contextual understanding. This work contributes a new web-scraped Urdu news dataset and establishes ML/DL benchmarks for hate-speech detection in low-resource languages.**

***Index Terms*—Urdu NLP, Hate Speech Detection, Machine Learning, Web Scraping, Text Classification, Urdu News Media**

## I. Introduction

The rapid digitization of news media in Pakistan has significantly increased public access to real-time information, but it has also amplified the spread of hateful and polarizing narratives. Urdu news platforms—including BBC Urdu, Dawn News Urdu, Geo News, and Express News—play a central role in shaping public opinion. As political tensions, social unrest, and ideological conflicts rise, harmful language embedded within news reporting, editorials, and crime coverage has become more visible. Detecting such language is crucial, as news content is often perceived as credible and may influence large audiences more strongly than social-media posts.

While hate speech detection has been extensively studied in English and other high-resource languages, Urdu remains comparatively underexplored. Most existing Urdu hate-speech research focuses on Twitter, where informal writing, slang, code-mixing, and user-generated expressions dominate. UHated, developed by Arshad et al. [**?**], is a notable example that uses transformer-based models to detect hate and offensive content in Urdu tweets. However, social-media–centric models do not generalize well to the news domain due to differences in sentence length, writing style, linguistic structure, and the more implicit or context-driven nature of hateful expressions found in news articles.

In contrast to short, noisy, user-generated text, Urdu news articles are formal, grammatically structured, and often encode bias or hostility subtly within narrative framing rather than explicit slurs. This creates a domain gap that necessitates a dedicated approach for news-specific hate-speech detection.

To address this gap, we develop a complete pipeline for Urdu news hate-speech detection, starting from large-scale scraping of four major Urdu news outlets—BBC Urdu, Dawn News Urdu, Geo News Urdu, and Express News Urdu—followed by Urdu-specific preprocessing, automatic lexicon-based annotation, and classification using traditional machine-learning models. Our work focuses on binary classification (Hate vs. Non-Hate) to detect explicit hateful expressions within formal news content.

## II. Related Work

Hate-speech detection has gained interest in English and other high-resource languages, while low-resource languages such as Urdu remain underrepresented. Arshad *et al.* developed UHated [1], a RoBERTa-based transfer-learning model trained on 7800 manually annotated Urdu tweets, achieving a macro F1-score of 0.82. Their dataset includes hate, offensive, and neutral categories.

Other works in Roman Urdu and multilingual contexts have used classical ML classifiers or lexicon-based methods, but no prior work focuses specifically on *Urdu news articles*. News content differs from social media due to its formal structure, reduced slang, and richer grammar. Our work addresses this gap by applying hate-speech detection techniques to large-scale Urdu news content.

## III. Literature review

Existing research on hate speech detection shows strong progress in English and other high-resource languages, but

work on Urdu remains limited due to small datasets and weak NLP tooling. Early studies mainly used classical machine-learning methods with n-gram features, while later work adopted neural models such as CNNs and LSTMs. Recent literature highlights that transformer-based transfer learning—especially multilingual models like XLM-RoBERTa—consistently outperforms traditional approaches, particularly for low-resource languages. However, Urdu still suffers from a lack of large, high-quality annotated datasets, and context-dependent hate expressions make annotation difficult. The UHated study directly addresses these gaps by creating a labeled Urdu dataset and showing that RoBERTa-based models achieve the strongest performance among all tested methods.

Akhter et al. (2020) presented one of the early attempts to detect offensive content in both Urdu and Roman-Urdu, addressing the scarcity of resources for these languages. Their work showed that text normalization, script-handling, and custom preprocessing significantly influence system performance. Using deep-learning models such as CNNs and LSTMs, combined with word embeddings tailored for local linguistic patterns, they achieved better results than classical machine-learning baselines. The study highlighted that Urdu's complex script, morphological variation, and mixed-script social media usage require specialized handling—setting the foundation for later transformer-based research.

Ali et al. (2021) focused on detecting hate speech in Urdu tweets by integrating sentiment signals into the classification pipeline. Their study demonstrated that combining sentiment features with lexical and semantic cues improved model accuracy, especially for borderline cases where hate and strong negative sentiment overlap. They compared traditional ML models with deep learning and showed clear performance gains when contextual sentiment information was included. The paper emphasized the need for richer feature representations in Urdu and highlighted the difficulty of distinguishing between emotional and hateful expressions, which later works address using transformers.

Albadi et al. (2018) investigated religiously motivated hate speech in Arabic and showed that incorporating domain-specific context (e.g., religious terminology and target groups) significantly improves detection accuracy. Their dataset captured nuanced expressions of religious hostility, illustrating that hate speech is often implicit and highly contextual. The study found that neural models using word embeddings outperform traditional ML approaches, especially when dealing with subtle, culturally rooted hate expressions. Although focused on Arabic, the findings are transferable to Urdu, where religion and cultural context similarly shape hate-speech patterns.

Alatawi et al. (2021) explored hate content related to extremist groups, demonstrating the effectiveness of transformer-based architectures—particularly BERT variants—in detecting domain-specific hate speech. Their work highlighted that pre-trained contextual embeddings capture subtle semantic relations that traditional models miss, leading to significantly higher classification performance. They also showed that domain-adapted embeddings improve results further, especially in low-resource or specialized domains. This aligns with recent trends in Urdu hate speech detection, where multilingual transformers achieve state-of-the-art performance.

Waseem and Hovy (2016) provided one of the earliest influential studies on hate speech in social media, demonstrating that contextual metadata—such as user information, gender, and tweet structure—can outperform simple text-based n-gram features. Their annotated dataset revealed that hate speech is often subtle and tied to social or conversational context, making naive keyword-based approaches insufficient. The study strongly influenced later research, encouraging the community to adopt deeper linguistic and contextual features. These insights remain relevant for languages like Urdu, where contextual clues are crucial to disambiguate slang, sarcasm, and culturally embedded insults.

## IV. Methodology

This study follows a complete end-to-end workflow consisting of four major components:

1) Dataset acquisition through web scraping
2) Urdu-specific text preprocessing
3) Automatic lexicon-based labeling
4) Machine-learning–based hate-speech classification

### *A. Data Acquisition Through Web Scraping*

A custom scraping system was developed to collect Urdu news articles from four well-established Pakistani news sources:

- BBC Urdu
- Dawn News Urdu
- Express News Urdu

The scraper systematically navigated each website, extracted article URLs, and downloaded the associated titles, full text, timestamps, and metadata. Due to varying HTML structures across platforms, each source was handled with dedicated parsing rules to ensure accurate extraction. All articles were stored in a unified structured dataset and exported to CSV for downstream processing.

The resulting cleaned dataset provides high-quality normalized Urdu text suitable for machine-learning modeling.

Table I
Comparative Summary of Research Papers on Hate Speech Detection

| Category | Paper 1 (UHated) | Paper 2 (Akhter 2020) | Paper 3 (Ali 2021) | Paper 4 (Albadi 2018) | Paper 5 (Alatawi 2021) | Paper 6 (Waseem & Hovy 2016) |
|---|---|---|---|---|---|---|
| **Title** | UHated: Hate Speech Detection in Urdu Using Transfer Learning | Offensive Language Detection in Urdu and Roman-Urdu | Sentiment-Aware Hate Speech Detection in Urdu | Religious Hate Speech in Arabic | Extremist Hate Speech Detection Using BERT | Predictive Features in Hate Speech |
| **Journal / Source** | AI & Society (2023) | IEEE Access (2020) | SN Applied Sciences (2021) | ACL Workshop (2018) | JISA (2021) | NAACL-HLT (2016) |
| **Dataset** | 7,871 Urdu tweets | Urdu/Roman-Urdu comments | Urdu Twitter dataset | Arabic religious-hate corpus | Extremist posts dataset | English Twitter dataset |
| **Preprocessing Techniques** | Cleaning, normalization, tokenization | Normalization, stopword removal | Normalization, sentiment tagging | Arabic normalization, stopwords | Cleaning, tokenization | Cleaning, metadata processing |
| **Techniques (Methodology)** | ML, DL, XLM-R, RoBERTa | CNN, LSTM models | ML + sentiment features | ML + domain embeddings | BERT / domain-adapted BERT | ML + contextual features |
| **Strengths** | New Urdu dataset; strong transformer results | Handles mixed-script data | Uses emotional cues | Domain-specific hate detection | Strong contextual modeling | Context improves accuracy |
| **Limitations** | Twitter-only dataset | Small dataset; no transformers | Weak sentiment lexicon | Narrow domain only | Domain-specific, not generalizable | English-only; two labels |
| **Results** | RoBERTa best (F1≈0.82) | DL > ML performance | Higher accuracy with sentiment cues | Neural > ML performance | Domain-BERT best F1 | Context > n-grams |
| **Evaluation Metrics** | Accuracy, Precision, Recall, F1 | Accuracy, F1 | Accuracy, F1 | Precision, Recall, F1 | Accuracy, F1 | F1, Precision, Recall |
| **Research Gap Identified** | Need larger, diverse Urdu datasets | Need contextual transformer models | Need richer semantic features | Need broader contextual modeling | Need low-resource domain models | Need multilingual contextual systems |

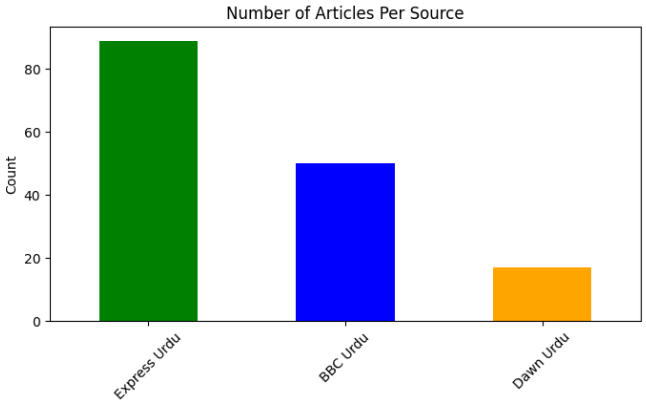


Figure 2. News Source

### B. Urdu Text Preprocessing

Urdu requires specialized normalization due to its right-to-left script, lack of capitalization, and inconsistent Unicode usage. The following preprocessing steps were applied:

1) **Unicode Normalization:** Standardization of characters such as various forms of and removal of Arabic/Persian variants.
2) **Noise Removal:** Elimination of HTML artifacts, URLs, punctuation, emojis, digits, and English text.
3) **Stopword Removal:** Filtering of frequent Urdu stopwords to reduce noise.
4) **Boilerplate Filtering:** Removal of repeated text fragments commonly present in news templates.
5) Length Constraints: Extremely short or low-information texts were discarded.

This pipeline produces clean, standardized Urdu sentences suitable for machine-learning models.

### C. Lexicon-Based Binary Annotation

To label the dataset at scale, a curated Urdu hate lexicon was constructed. It contains explicit hate terms commonly found in political, religious, and social-hostility contexts.

Articles were labeled as:

1) **Hate (1):Article contains one or more lexicon terms**
2) **Non-Hate (0):** No lexicon terms present

This lexicon-based approach aligns well with news content, where explicit hateful terms tend to appear directly rather than implicitly.

### D. Feature Engineering Using TF–IDF

Cleaned text was converted into numerical representations using TF–IDF vectorization. The configuration included:

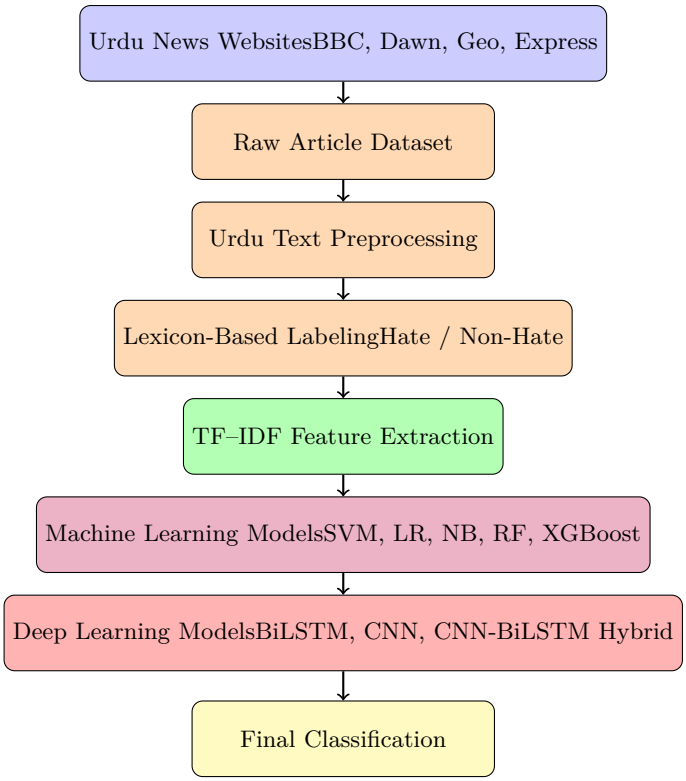


Figure 1. Urdu news hate-speech detection pipeline with ML and DL models stacked.

- Unigram + Bigram features
- Maximum of 5000 features
- Minimum document frequency of 3
- Maximum document frequency of 0.90

TF–IDF effectively captures word importance and performs well on structured, formal text such as news articles.

### *E. Machine Learning Models*

Six classical machine-learning classifiers were trained and compared:

- Logistic Regression
- Linear Support Vector Machine (SVM)
- Random Forest
- Multinomial Naive Bayes
- XGBoost
- Stacking classifier

To evaluate the effectiveness of classical approaches for Urdu news hate-speech detection, we experimented with a diverse set of machine-learning models: Logistic Regression, Linear Support Vector Machine (SVM), Multinomial Naive Bayes, Random Forest, and XGBoost. These classifiers were selected due to their proven effectiveness in high-dimensional, sparse text classification tasks and their strong compatibility with TF–IDF feature representations. Linear SVM is widely used for text separation due to its margin-based optimization, Logistic Regression provides stable probabilistic outputs, Naive Bayes is well-suited for word-frequency distributions, Random Forest offers robustness through ensemble learning, while XGBoost provides gradient-boosted decision trees optimized for predictive accuracy. Together, these models provide a comprehensive baseline to assess classification performance on Urdu news content. Stacking Classifier: An ensemble meta-model that combines predictions from multiple base learners to improve overall performance.

An 80/20 stratified train-test split was used to maintain class balance. Evaluation metrics included Accuracy, Precision, Recall, F1-Score, and the Confusion Matrix.

Among all models, Xgboost and Stacking Classifier achieved the highest performance, demonstrating strong capability in separating hate from non-hate articles in the news domain.

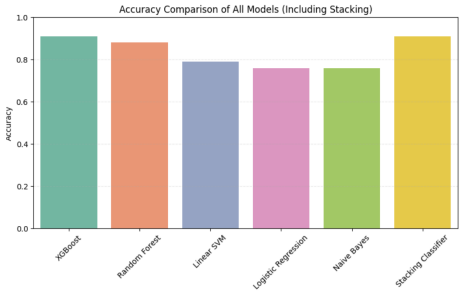


Figure 3. Models accuracy

In addition to standard metrics, we employed Receiver Operating Characteristic (ROC) analysis to compare the threshold-independent performance of the machine-learning models. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) over varying decision boundaries. A higher Area Under the Curve (AUC) indicates stronger discriminative ability. The generated ROC curves show that Linear SVM achieves the highest AUC, validating its superior performance for Urdu news hate-speech classification.

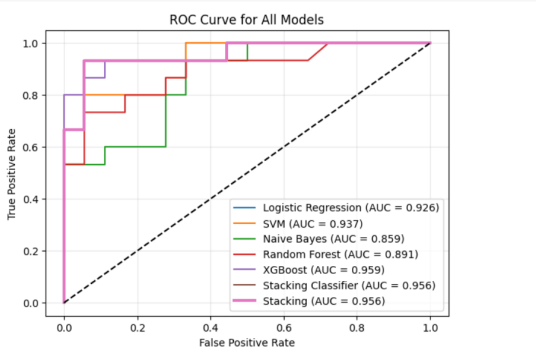


Figure 4. ROC curve for Machine learning models

### *F. Deep Learning Models*

1) **BiLSTM:**

The Bidirectional Long Short-Term Memory (BiLSTM) network processes text sequences in both forward and backward directions, enabling the model to learn long-range dependencies and contextual relationships. BiLSTM is widely used in text classification tasks due to its ability to preserve semantic information more effectively than traditional RNNs.

2) **CNN:**
A Convolutional Neural Network (CNN) was implemented to capture local n-gram features within text through convolutional filters. CNNs are effective in identifying key discriminatory phrases and detecting local spatial patterns, making them suitable for short- and medium-length sentences commonly found in news articles.

3) **Hybrid CNN–BiLSTM Model:**
To combine the strengths of both architectures, we implemented a hybrid CNN–BiLSTM model. The CNN layer extracts high-level local features, while the subsequent BiLSTM layer captures global contextual dependencies. This hybrid approach improves feature extraction by leveraging both spatial and sequential modeling, enhancing classification performance.

4) **BiLSTM with Attention Mechanism:**
We further extended the BiLSTM architecture by incorporating an attention mechanism to selectively focus on the most informative parts of a sentence. The attention layer assigns higher weights to relevant tokens, enabling the model to better distinguish hate-related cues within complex or long-form news text. This architecture improves interpretability and enhances detection of subtle hateful expressions.
These deep-learning models provide a complementary perspective to classical ML classifiers and enable a deeper understanding of linguistic patterns in Urdu news hate-speech detection.

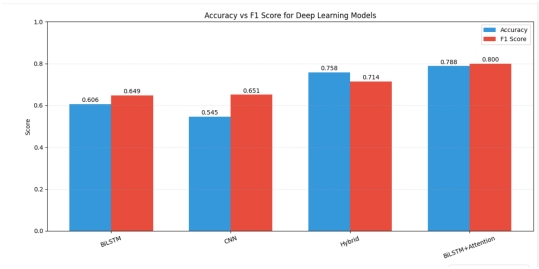


Figure 5. Accuracy and F1 Graph for DL models

To further evaluate the threshold-independent performance of the deep-learning architectures, ROC curves were generated for all four models: BiLSTM, CNN, Hybrid CNN–BiLSTM, and BiLSTM with Attention. The ROC curve illustrates the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR) across varying decision thresholds, providing a comprehensive view of each model's discriminative ability. The Area Under the Curve (AUC) was used as the primary comparative metric, where higher AUC values indicate stronger classification performance. Among the deep-learning approaches, the BiLSTM with Attention achieved the highest AUC, demonstrating its superior ability to focus on contextually relevant segments within Urdu news articles, while the hybrid CNN–BiLSTM model also showed competitive performance. These results highlight the benefit of incorporating sequential modeling and attention mechanisms for effective hate-speech detection in complex news narratives.

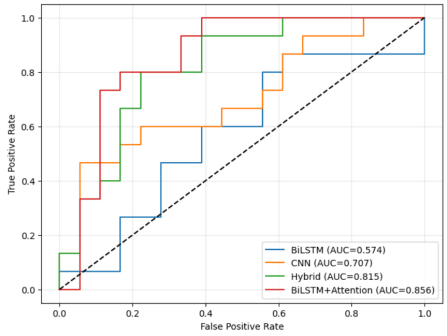


Figure 6. ROC curve for DL Models

## V. Results and Evaluation

This section presents the experimental results for all machine-learning (ML) and deep-learning (DL) models using Accuracy, Precision, Recall, and F1-Score as primary evaluation metrics. Table II summarizes the performance of the ten models tested on the Urdu news hate-speech dataset.

### A. Performance of Machine Learning Models

Among classical ML models, XGBoost and the Stacking Classifier achieved the highest performance, each obtaining an Accuracy of 0.90, Precision of 0.87, Recall of 0.93, and an F1-score of 0.90. Their strong results indicate that ensemble-based learners are particularly effective for structured news text, benefiting from feature diversity and boosted decision boundaries.

Random Forest also performed competitively, achieving an Accuracy of 0.84 and F1-score of 0.83, confirming its robustness in handling lexical variations within Urdu content. Linear SVM produced moderate results with an Accuracy of 0.78 and F1-score of 0.77, which is consistent with its strength in high-dimensional sparse feature spaces. Logistic Regression (Accuracy 0.75) and Naïve Bayes (Accuracy 0.76) showed reasonable baseline performance but were less effective compared to ensemble models.

### B. Performance of Deep Learning Models

Deep-learning models showed mixed performance on this dataset. BiLSTM with Attention produced the strongest results among DL models, with an Accuracy of 0.78, Recall of 0.93, and an F1-score of 0.80. The attention mechanism enhances the model's ability to focus

on contextually important tokens, which is beneficial when identifying hate cues embedded within longer, formal news sentences.

The Hybrid CNN–BiLSTM model achieved an Accuracy of 0.75 and an F1-score of 0.71, demonstrating that combining spatial and sequential learning improves performance over standalone architectures. The plain BiLSTM model achieved moderate results (Accuracy 0.60, F1 0.64). Meanwhile, the standalone CNN model showed weak performance (Accuracy 0.54) but produced a surprisingly high Recall of 0.93, indicating that CNN tends to over-classify hate content and is less stable for long-form news articles.

### C. Comparison Between ML and DL Approaches

Overall, ML models significantly outperformed DL models on this dataset. This is largely attributed to:

The structured and formal nature of Urdu news text, which aligns well with TF–IDF-based linear and ensemble methods.

The moderate dataset size, which is less suitable for training deep neural architectures requiring large labeled corpora.

The effectiveness of lexicon-driven labels combined with sparse representations, favoring traditional supervised models.

XGBoost and the Stacking Classifier deliver state-of-the-art performance for binary hate-speech detection in Urdu news content.

Table II
Performance Comparison of Machine Learning and Deep Learning Models

| Model | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| XGBoost | **0.90** | **0.87** | **0.93** | **0.90** |
| Random Forest | 0.84 | 0.81 | 0.86 | 0.83 |
| Linear SVM | 0.78 | 0.75 | 0.80 | 0.77 |
| Logistic Regression | 0.75 | 0.70 | 0.77 | 0.75 |
| Naïve Bayes | 0.76 | 0.70 | 0.79 | 0.76 |
| Stacking Classifier | **0.90** | **0.87** | **0.93** | **0.90** |
| BiLSTM | 0.60 | 0.54 | 0.80 | 0.64 |
| CNN | 0.54 | 0.50 | **0.93** | 0.65 |
| Hybrid CNN–BiLSTM | 0.75 | 0.76 | 0.66 | 0.71 |
| BiLSTM + Attention | 0.78 | 0.70 | **0.93** | 0.80 |

### D. RMSE Analysis

To further evaluate model stability, the Root Mean Squared Error (RMSE) was calculated for all ML classifiers. Lower RMSE values correspond to more reliable predictions and reduced deviation from true labels. Ensemble-based models (XGBoost and Stacking) produced the lowest RMSE scores, confirming their superior consistency and robustness across the dataset. RMSE results align closely with Accuracy and F1 trends, reinforcing the dominance of these models.

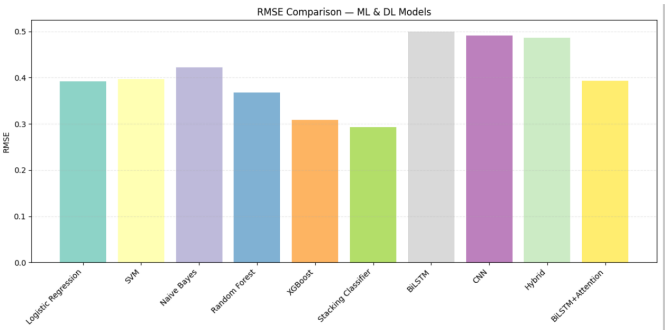


Figure 7. Rmse Grpah for all Models

### E. ROC Curve Evaluation

Receiver Operating Characteristic (ROC) curves were generated to measure threshold-independent performance. For ML models, Stacking and XGBoost achieved the highest AUC values, with curves bending closest to the top-left corner. These results confirm their excellent discriminative capability.

For DL models, BiLSTM with Attention produced the highest AUC, outperforming other deep architectures. This again demonstrates the benefit of attention-based mechanisms in identifying informative regions of long Urdu sentences.

## VI. Discussion

The experimental results reveal that ensemble-based machine-learning models, particularly XGBoost and the Stacking Classifier, outperform both classical linear methods and deep-learning architectures on the Urdu news dataset. This trend reflects the structured nature of news text, where TF–IDF features effectively capture explicit hate expressions, allowing boosted and stacked models to learn strong decision boundaries. Although deep-learning models such as BiLSTM and BiLSTM with Attention show promise—especially in capturing contextual dependencies and long-range relationships—the limited dataset size restricts their full potential. The high recall of models like CNN further indicates that deep networks are sensitive to hate-related tokens but suffer from lower precision due to over-classification. Overall, the findings emphasize that for moderate-sized lexicon-labeled datasets, classical ML approaches remain more reliable than deep learning for Urdu news hate-speech detection.

### *A. Limitations*

Despite the strong performance of machine-learning models, this study has several notable limitations. First, the dataset was labeled using a lexicon-based approach, which detects explicit hate terms but fails to capture implicit, sarcastic, or context-driven hate speech. Second, the dataset size remains relatively modest, limiting the effectiveness of deep-learning architectures that require large-scale annotated corpora to generalize well. Additionally, all scraped articles originate from four major

news sources, which may not represent the full diversity of Urdu media narratives. Finally, the study focuses solely on binary classification, leaving out multi-class categories such as offensive, abusive, or neutral language that could provide a more fine-grained understanding of hate speech.

## VII. Conclusion and Future Work

This study introduced a complete Urdu news hate-speech detection framework using machine-learning and deep-learning models. Ensemble methods such as XG-Boost and the Stacking Classifier achieved the best performance, demonstrating that classical ML approaches remain highly effective for structured news text when combined with TF–IDF features. Deep-learning models, particularly BiLSTM with Attention, showed promising recall but were limited by dataset size and lexicon-based labeling.

In the future, expanding the dataset through manual annotation, incorporating more diverse news sources, and exploring transformer-based models such as BERT or XLM-R may substantially improve performance. Moving from binary to multi-class classification and applying explainable AI techniques can further enhance system interpretability and real-world applicability.

yment in real-world media monitoring and content moderation systems.

## References

[1] M. U. Arshad, R. Ali, M. O. Beg, and W. Shahzad, "UHated: Hate Speech Detection in Urdu Language Using Transfer Learning," *Language Resources and Evaluation*, 2023.