An Explainable Ensemble Learning Framework for Accurate Fake News Detection Using TF-IDF Features

Pooja Ashok Wagh[1]
[1]Research Scholar, CSE Department, Oriental University, Indore
Dr. Akanksha Pal[2]
[2]Asst. Professor, CSE Department, Oriental University, Indore
akankshapal@orientaluniversity.in

Abstract—The rapid proliferation of misinformation on digital platforms has become a critical societal challenge. This paper presents a comprehensive comparative study of four gradient boosting ensemble classifiers — Gradient Boosting Classifier (GBC), XGBoost, LightGBM, and CatBoost — for automated fake news detection. The publicly available Kaggle Fake and Real News Dataset is used as the benchmark. A systematic text preprocessing pipeline is applied, followed by Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction using unigrams and bigrams. All four models are trained on an 80/20 stratified split and evaluated using Accuracy, Precision, Recall, F1-Score, ROC-AUC, and RMSE. LightGBM achieves the highest F1-Score of 0.9959 and ROC-AUC of 0.9996, followed closely by Gradient Boosting (F1=0.9958), XGBoost (F1=0.9956), and CatBoost (F1=0.9955). LIME (Local Interpretable Model-Agnostic Explanations) analysis is applied to the best model to identify the most influential textual features driving predictions. Error analysis through misclassification counts further validates the robustness of all classifiers. The results confirm that TF-IDF-based gradient boosting ensembles provide highly accurate, efficient, and interpretable solutions for real-world fake news detection.

Keywords—fake news detection; gradient boosting; XGBoost; LightGBM; CatBoost; TF-IDF; natural language processing; LIME; misinformation.

I. INTRODUCTION

The proliferation of social media platforms has dramatically accelerated the speed at which information, and misinformation, spreads. Fake news — fabricated or deliberately misleading content presented as factual journalism — has emerged as a significant threat to democratic processes, public health, and social stability. Manual fact-checking is neither scalable nor fast enough to counter this phenomenon, making automated detection systems a research priority.

Natural Language Processing (NLP) and machine learning (ML) offer powerful tools for text-based classification. Among the most effective ML approaches for structured and semi-structured data are gradient boosting ensembles, which iteratively combine weak learners to build strong predictors. When paired with TF-IDF, a robust statistical text representation, these models can capture discriminative linguistic patterns that differentiate fake from real news.

This paper makes the following specific contributions: (1) a reproducible end-to-end fake news detection pipeline on the Kaggle benchmark dataset; (2) a rigorous comparative study of four gradient boosting classifiers — GBC, XGBoost, LightGBM, and CatBoost — under identical experimental conditions; (3) comprehensive exploratory data analysis (EDA) including label distribution, text length analysis, word clouds, and n-gram analysis; (4) LIME-based model

interpretability to explain individual predictions; and (5) error analysis through misclassification counts to understand model behavior.

The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the dataset. Section IV presents the proposed methodology and workflow. Section V covers exploratory data analysis. Section VI reports experimental results. Section VII analyzes model interpretability using LIME. Section VIII discusses findings and limitations. Section IX concludes the paper.

II. RELATED WORK

Fake news detection has been studied extensively from multiple perspectives. Shu et al. [1] surveyed the field and categorized approaches into knowledge-based, style-based, propagation-based, and source-credibility methods. Style-based approaches, which analyze writing style and linguistic cues, are particularly relevant to text-only classification pipelines such as the one proposed here.

Ahmed et al. [2] demonstrated that combining n-gram features with linear classifiers — Logistic Regression and Linear SVM — achieved over 92% accuracy on news classification tasks. Their work established n-grams as highly effective discriminative features for fake news detection. Granik and Mesyura [3] showed that even a simple Naive Bayes classifier with bag-of-words features could achieve reasonable detection performance.

Gradient boosting methods have demonstrated superior performance across classification tasks. Chen and Guestrin [4] introduced XGBoost, achieving state-of-the-art results with regularized gradient boosting. Ke et al. [5] proposed LightGBM, which uses histogram-based leaf-wise tree growth for faster training while maintaining competitive accuracy. Prokhorenkova et al. [6] proposed CatBoost with ordered boosting, reducing prediction bias. Despite extensive research, a unified comparison of all four gradient boosting algorithms under identical TF-IDF preprocessing and dataset conditions in the fake news domain is underexplored — motivating the present study.

Regarding interpretability, Ribeiro et al. [7] proposed LIME to explain individual model predictions by fitting local surrogate models. LIME has been applied to NLP tasks to identify which words most influence a classifier's decision, improving trust and transparency in deployed systems.

III. DATASET

The Kaggle Fake and Real News Dataset [8] is used in this study. It contains two CSV files: Fake.csv with 23,481 fabricated news articles and True.csv with 21,417 authentic news articles, totalling 44,898 samples. Each record contains the article title, body text, subject category, and publication date. Labels are assigned as 0 (Fake) and 1 (Real). The combined dataset is randomly shuffled with a fixed seed (random_state=42) to ensure reproducibility. The article title and body text are concatenated into a single content field for classification.

A. Label Distribution

Fig. 1 shows the class distribution. Fake news articles (label=0) total 23,481 and real news articles (label=1) total 21,417, yielding a near-balanced distribution. This approximate balance minimizes class-imbalance bias during model training and ensures reliable metric computation without requiring oversampling or undersampling strategies.

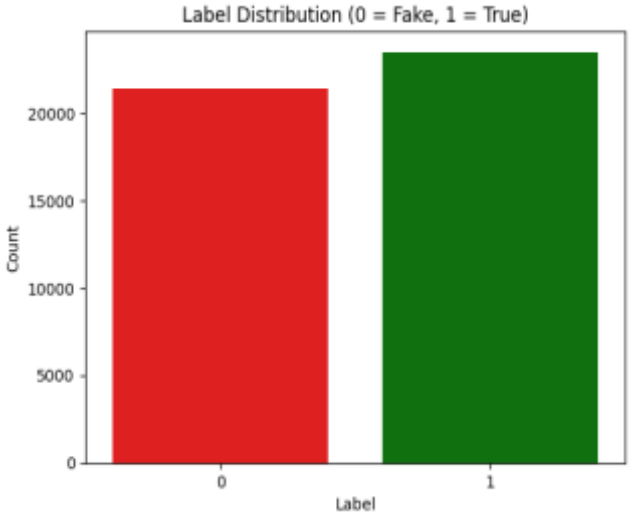

Fig. 1. Label Distribution of the Dataset (0 = Fake, 1 = Real).

IV. PROPOSED METHODOLOGY

A. System Workflow

Fig. 2 illustrates the complete end-to-end pipeline of the proposed fake news detection system, from raw data collection through preprocessing, feature extraction, model training, evaluation, and interpretability analysis.

Step 1: Data Collection — Kaggle Fake & Real News dataset (44,898 articles; Fake.csv + True.csv)

▼

Step 2: Labeling — Assign label=0 (Fake) and label=1 (Real); merge into single DataFrame

▼

Step 3: Text Preprocessing — Lowercase → Remove URLs/HTML/mentions → Remove punctuation → Tokenize → Stopword removal → Remove short words → Stemming (optional) → Rejoin tokens

▼

Step 4: Feature Engineering — TF-IDF Vectorization (max_features=5000, ngram_range=(1,2))

▼

Step 5: Train / Test Split — 80% training (35,918 samples) | 20% testing (8,980 samples), stratified

▼

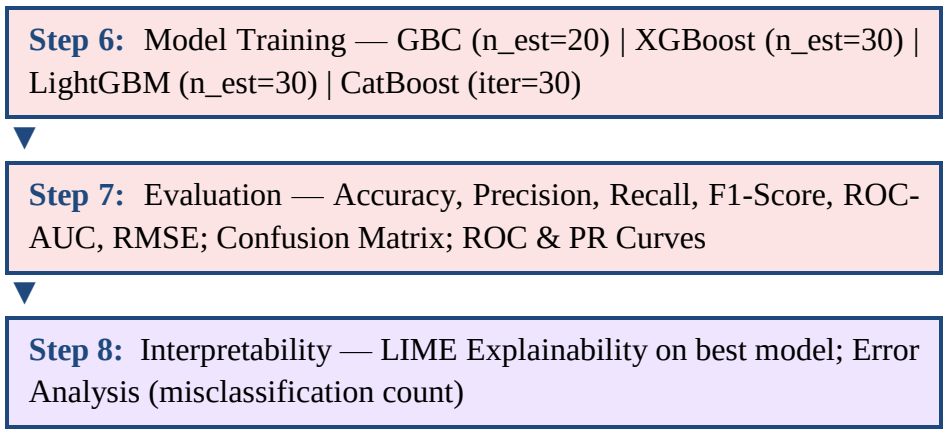

Fig. 2. Proposed end-to-end pipeline for fake news detection.

B. Text Preprocessing

Raw news article text is processed through a multi-step preprocessing pipeline to standardize content and remove noise before feature extraction. The steps are applied in the following sequence:

(1) Lowercasing: All characters are converted to lowercase to ensure case-insensitive vocabulary (e.g., "Hello" → "hello").

(2) URL Removal: Web links (http://, www.) are deleted as they carry no semantic meaning for classification.

(3) HTML Tag Removal: Markup tags (<p>,
,) are stripped to retain only visible text content.

(4) Mention and Hashtag Removal: Social media tokens (@username, #topic) are removed as they introduce noise.

(5) Punctuation, Digit, and Special Character Removal: Only alphabetic characters (a–z) and spaces are retained.

(6) Tokenization: The cleaned text is split into individual word tokens.

(7) Stopword Removal: Common English function words (e.g., "the", "is", "and") are removed using the NLTK stopwords corpus.

(8) Short Word Removal: Single-character tokens are discarded, as they generally carry no meaningful information.

(9) Stemming (Optional): Porter Stemmer is applied to reduce words to their root form (e.g., "running" → "run").

(10) Token Rejoining: Processed tokens are rejoined into a cleaned sentence for downstream feature extraction.

C. Feature Extraction: TF-IDF Vectorization

Cleaned text is converted into numerical feature vectors using the TF-IDF (Term Frequency-Inverse Document Frequency) method. TF-IDF assigns higher weights to terms that appear frequently in a document but rarely across the corpus, thereby capturing discriminative information. The vectorizer is configured with: max_features=5,000 (top vocabulary terms by TF-IDF score), stop_words="english" (additional stopword filtering), and ngram_range=(1,2) to capture both unigrams and bigrams. The TF-IDF vectorizer is fitted exclusively on the training set and applied to both training and test sets to prevent data leakage.

D. Train/Test Split

The dataset is partitioned using stratified random sampling with an 80% training / 20% testing ratio (random_state=42). Stratification ensures that the original class proportions (approximately 52% Fake, 48% Real) are preserved in both splits. This yields 35,918 training samples and 8,980 test samples.

V. EXPLORATORY DATA ANALYSIS

Comprehensive exploratory analysis was performed to understand the linguistic properties and statistical characteristics of the dataset prior to model training.

A. Text Length Distribution

Fig. 3 shows histograms of raw and cleaned word counts for fake and real news articles. Real news articles tend to have a broader word count distribution with a concentration around 200–400 words after cleaning, reflecting detailed journalistic reporting. Fake news articles show greater variability, with many shorter articles concentrated below 100 words after cleaning. After preprocessing, the distributions become more compact in both classes, confirming the effectiveness of the text cleaning pipeline in reducing noise.

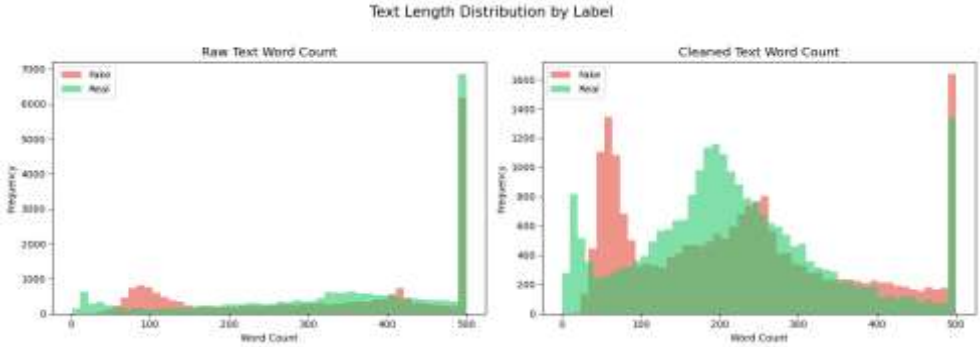

Fig. 3. Text Length Distribution by Label — Raw Text Word Count (left) and Cleaned Text Word Count (right).

B. Correlation Heatmap

Fig. 4 presents the Pearson correlation heatmap among numerical features: label, text_length, and clean_text_length. A near-perfect correlation (0.99) is observed between raw and cleaned text lengths, confirming that preprocessing preserves relative article length. Both length features show very low correlation with the label (0.06 and 0.01 respectively), indicating that article length alone is insufficient to distinguish fake from real news, and that richer textual features — such as TF-IDF — are necessary for effective classification.

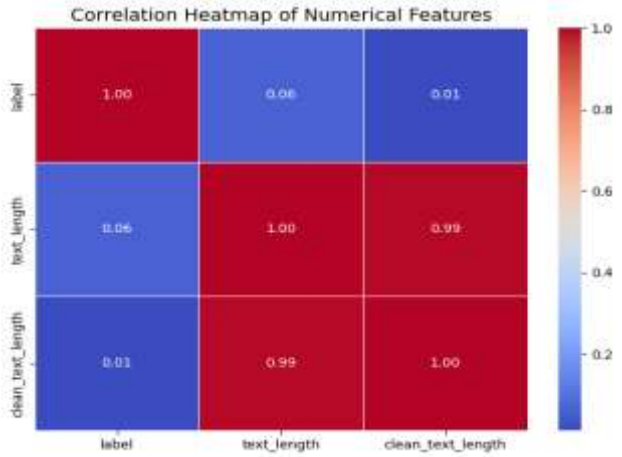

Fig. 4. Correlation Heatmap of Numerical Features (label, text_length, clean_text_length).

C. Word Cloud Visualization

Fig. 5 compares the most frequently occurring terms in fake news (Reds colormap) and real news (Greens colormap) articles. The word clouds are generated from the cleaned content of each class. Fake news is dominated by politically charged terms such as "said," "trump," "reuters," "state," "government," and "campaign," along with sensational vocabulary. Real news shows more neutral, journalistic language including institutional names ("president," "trump," "obama," "clinton"), policy terminology, and formal references. The size of each word in the visualization represents its frequency. These contrasting vocabulary patterns validate the discriminative potential of word-level TF-IDF features.

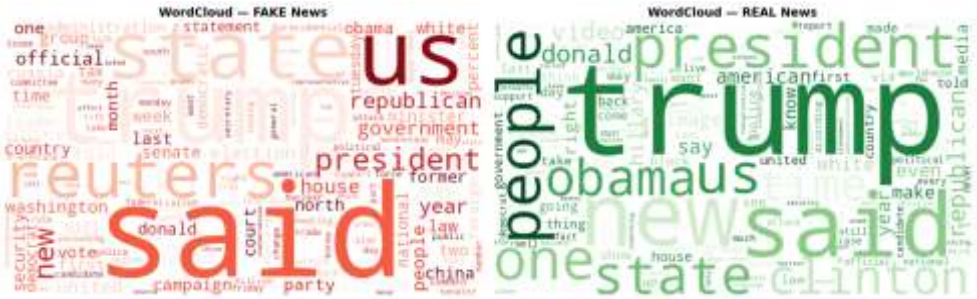

Fig. 5. Word Cloud Comparison: Fake News (left, Reds) vs. Real News (right, Greens).

D. N-gram Analysis

Fig. 6 presents the top 15 unigrams, bigrams, and trigrams for fake and real news articles. Fake news unigrams include "said," "trump," "us," and "president." Real news unigrams are led by "trump," "said," "president," and "people." At the bigram level, fake news prominently features "united states," "white house," and "donald trump," while real news bigrams include "donald trump," "hillary clinton," and "white house." Trigrams reveal deeper patterns: fake news contains "president donald trump," "washington reuters us," and "us president donald," while real news features "st century wire," "new york times," and "black lives matter." These n-gram patterns confirm that while both classes share some vocabulary, the combination and context of terms differ significantly, which TF-IDF bigrams effectively capture.

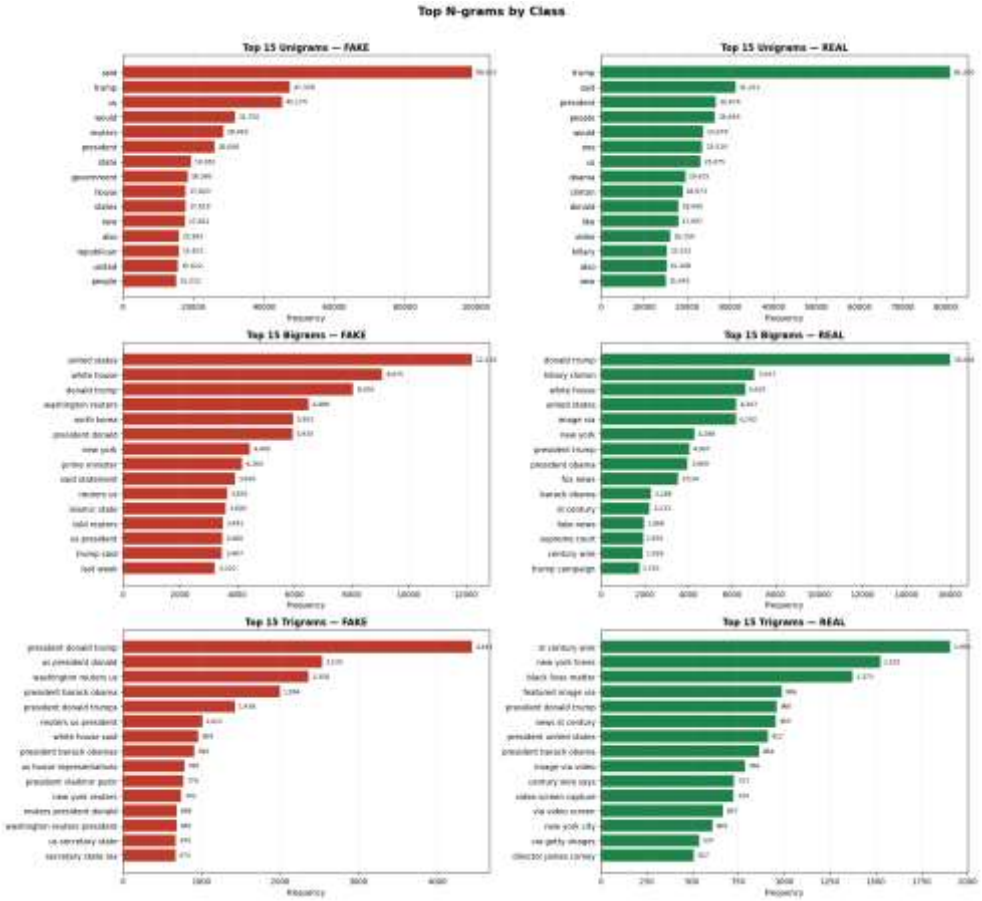

Fig. 6. Top 15 Unigrams, Bigrams, and Trigrams for Fake and Real News Articles.

VI. EXPERIMENTAL RESULTS

A. Quantitative Performance Summary

Table V presents the complete performance comparison of all four gradient boosting classifiers. All models achieve accuracy exceeding 99.5% on the 8,980-sample test set. LightGBM achieves the highest F1-Score (0.9959) and ROC-AUC (0.9996), indicating the best balance between precision and recall. Gradient Boosting achieves the highest Precision (0.9979) with the second-best F1-Score (0.9958). XGBoost matches LightGBM in ROC-AUC (0.9995) with slightly lower F1 (0.9956). CatBoost achieves the lowest F1 (0.9955) and highest RMSE (0.0684) among the four models, though still demonstrating strong performance. The extremely narrow performance gap across models (F1 range: 0.0004) confirms that TF-IDF features are highly effective for this task regardless of the specific boosting algorithm used.

TABLE V. Fake News Detection — Final Results Summary

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	RMSE
LightGBM	**0.9958**	**0.9977**	**0.9943**	**0.9959**	**0.9996**	**0.0651**
Gradient Boosting	0.9957	0.9979	0.9938	0.9958	0.9971	0.0659
XGBoost	0.9954	0.9976	0.9936	0.9956	0.9995	0.0676

CatBoost	0.9953	0.9976	0.9934	0.9955	0.9986	0.0684

B. Confusion Matrices

Fig. 7 presents the confusion matrices for all four models. In each matrix, the diagonal entries (True Negatives: correctly classified Fake articles; True Positives: correctly classified Real articles) overwhelmingly dominate. Gradient Boosting correctly classifies 4,274 fake and 4,667 real articles, with only 10 false positives and 29 false negatives. LightGBM achieves the fewest total false negatives (27), while GBC achieves the fewest false positives (10). The consistently high diagonal values across all models confirm that the classifiers are highly reliable for both classes.

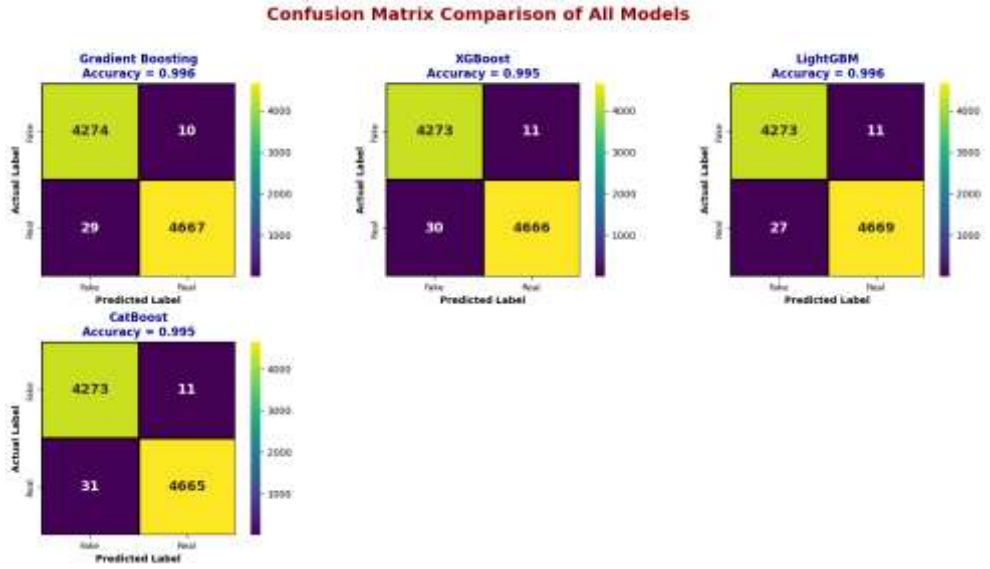

Fig. 7. Confusion Matrix Comparison for All Four Gradient Boosting Models.

C. ROC-AUC Curves

Fig. 8 shows the ROC curves (left) and ROC-AUC bar comparison (right) for all models. All curves hug the upper-left corner of the plot, confirming simultaneous high sensitivity and specificity. LightGBM and XGBoost achieve near-perfect AUC of 0.9996 and 0.9995 respectively. CatBoost (0.9986) and Gradient Boosting (0.9971) also perform excellently. The ROC analysis confirms that all four classifiers provide highly reliable probabilistic outputs suitable for threshold-tunable deployment scenarios.

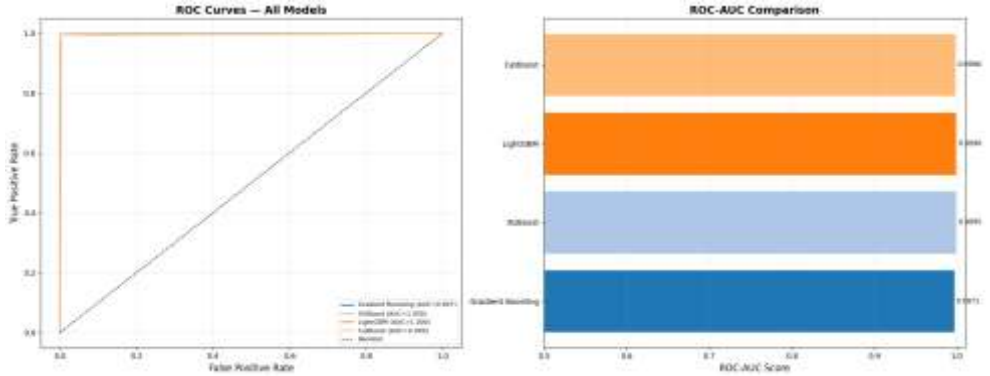

Fig. 8. ROC-AUC Curves (left) and ROC-AUC Score Comparison (right) for All Models.

D. Precision-Recall Curves

Fig. 9 presents the Precision-Recall (PR) curves for all four models. All models maintain near-perfect precision across the full range of recall values (Average Precision ≈ 1.000 for XGBoost and LightGBM; 0.999 for CatBoost; 0.996 for Gradient Boosting). Precision begins to drop only at extreme recall values close to 1.0, confirming that the models remain accurate even when tuned to detect the maximum number of fake news articles. This behavior is particularly important for real-world deployments where high recall is critical to minimize missed detections.

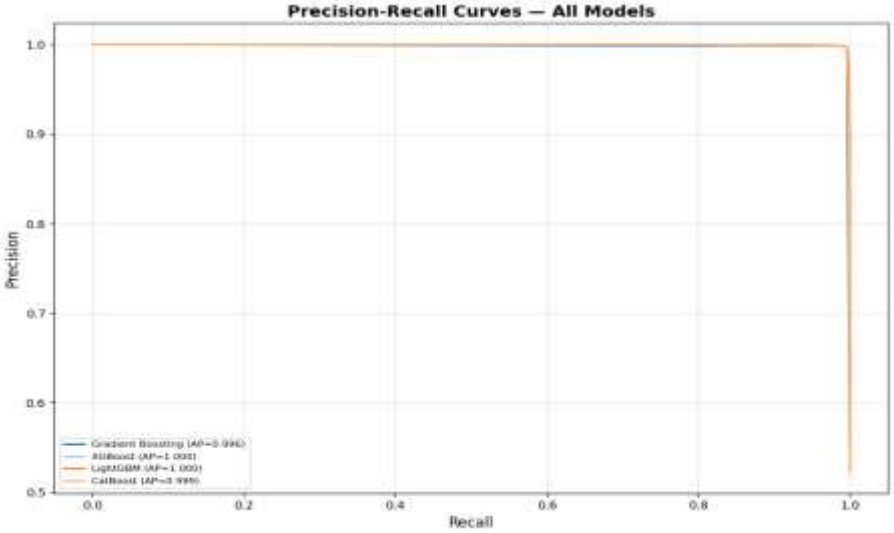

Fig. 9. Precision-Recall Curves for All Four Models.

E. Performance Metrics Comparison Dashboard

Fig. 10 presents a grouped bar chart comparing Accuracy, Precision, Recall, F1-Score, and ROC-AUC side-by-side for all four models on a shared scale (0.5–1.0). All bars approach 1.0 for every metric, confirming uniformly excellent performance. Minor differences are visible between models on the Recall and ROC-AUC bars. LightGBM and XGBoost show marginally higher bars overall, consistent with their leading F1 and AUC scores reported in Table V.

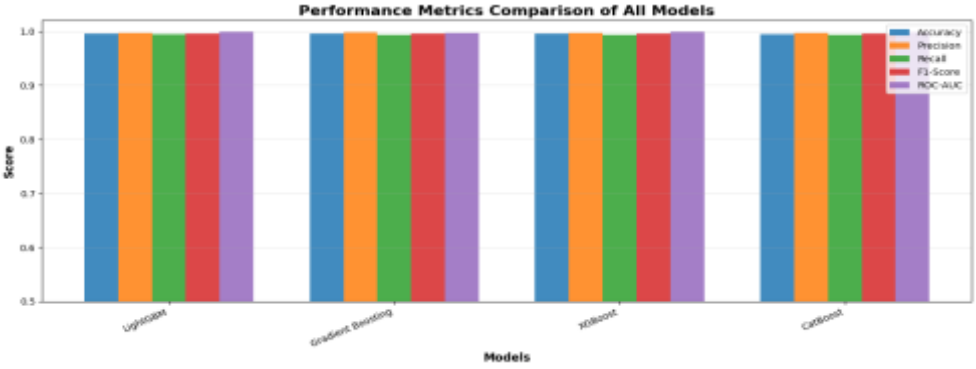

Fig. 10. Performance Metrics Comparison Dashboard — All Models and All Metrics.

F. Model Evaluation Heatmap

Fig. 11 provides a heatmap visualization of all evaluation metrics across models. The colormap highlights relative performance differences. LightGBM's yellow (bright) cell on the ROC-AUC column (1.000) stands out, as does XGBoost's matching score. Recall is the most differentiated metric, with CatBoost scoring 0.993 versus LightGBM's 0.994. The heatmap clearly confirms the marginal but consistent advantage of LightGBM over the other three models.

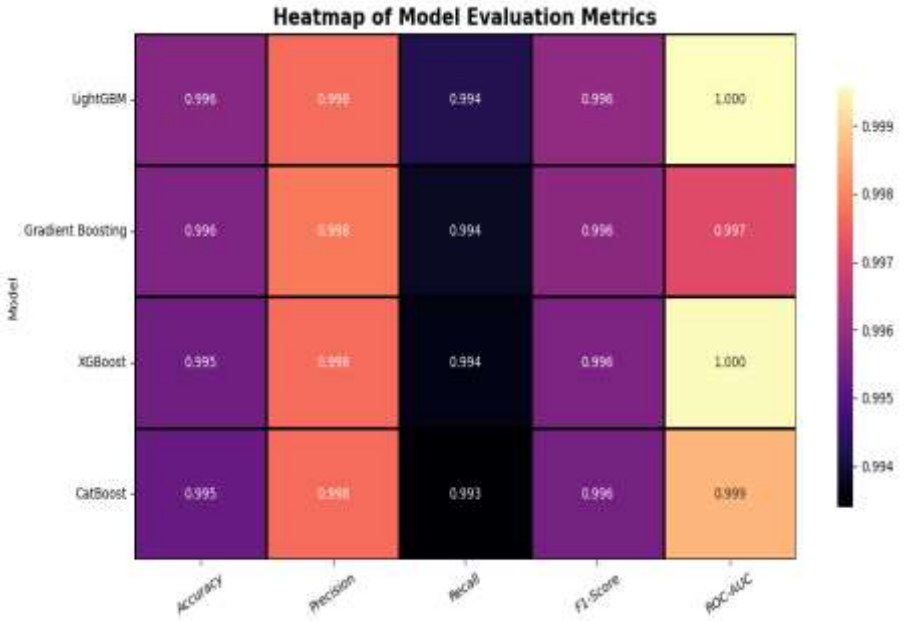

Fig. 11. Heatmap of Model Evaluation Metrics (Accuracy, Precision, Recall, F1-Score, ROC-AUC).

G. Radar Chart — Multi-metric Overview

Fig. 12 presents a radar chart overlaying all four models across five evaluation axes: Accuracy, Precision, Recall, F1-Score, and ROC-AUC. All four models occupy nearly the same polygon on the radar, filling out close to the outermost ring. This visualization confirms the near-equivalent, near-perfect multi-metric performance of the entire gradient boosting family on the Kaggle fake news benchmark when paired with TF-IDF features.

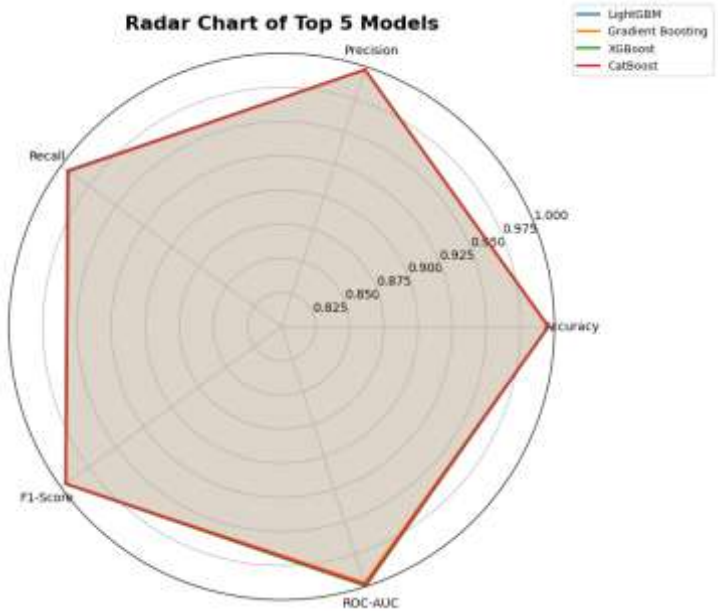

Fig. 12. Radar Chart — Multi-metric Performance Comparison of All Four Models.

H. F1-Score Ranking

Fig. 13 shows a horizontal bar chart ranking all models by F1-Score. LightGBM leads at 0.9959, followed by Gradient Boosting (0.9958), XGBoost (0.9956), and CatBoost (0.9955). The

ranking confirms LightGBM as the best-performing model, though the absolute difference across models is only 0.0004, demonstrating the robustness of TF-IDF features across different boosting algorithms.

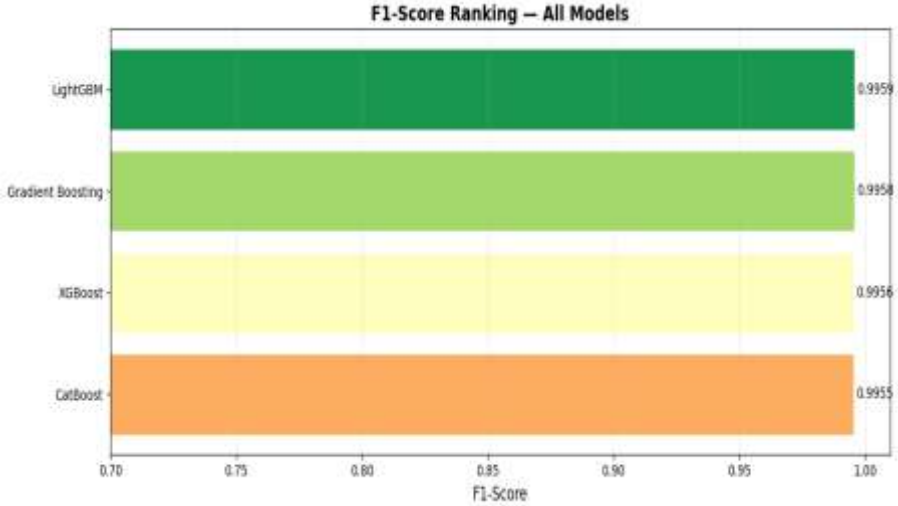

Fig. 13. F1-Score Ranking — All Gradient Boosting Models.

I. Calibration Curves

Fig. 14 presents the calibration curves of all four models against the perfectly calibrated reference line (dashed). A well-calibrated classifier produces probability estimates that reliably reflect the true likelihood of class membership. All four models show deviations from perfect calibration — a common characteristic of gradient boosting classifiers, which tend to produce overconfident probability estimates. Future work can address this using post-hoc calibration techniques such as Platt scaling (sigmoid calibration) or isotonic regression to improve probability reliability for downstream decision-making.

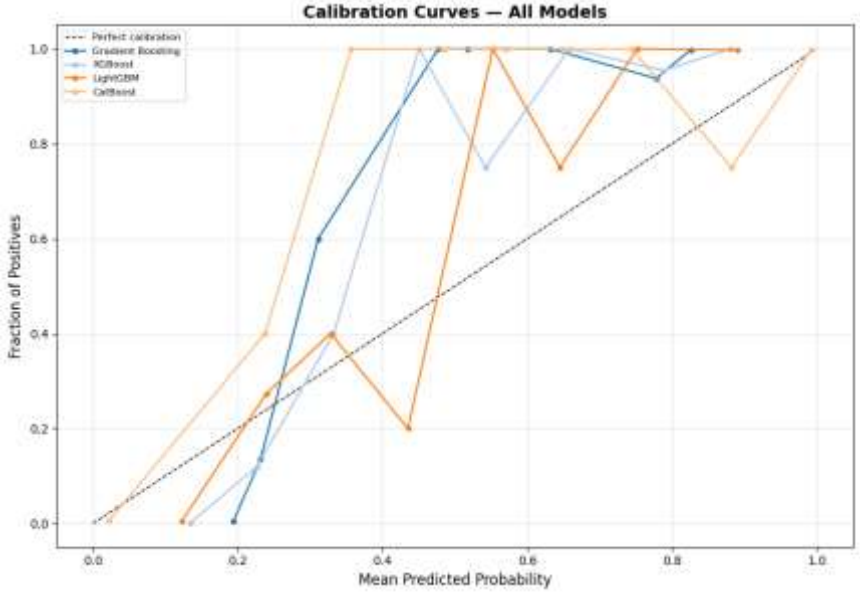

Fig. 14. Calibration Curves for All Four Models vs. Perfect Calibration Reference.

J. Error Analysis: Misclassification Count

Fig. 15 presents the misclassification count for each model. LightGBM produces the fewest misclassified samples (38), followed by Gradient Boosting (39), XGBoost (41), and CatBoost (42). The absolute differences are minimal, with only 4 samples separating the best and worst models. These low error counts on 8,980 test samples confirm the robustness of all four classifiers. Models with fewer misclassifications achieve higher classification accuracy and lower RMSE values, consistent with the quantitative summary in Table V.

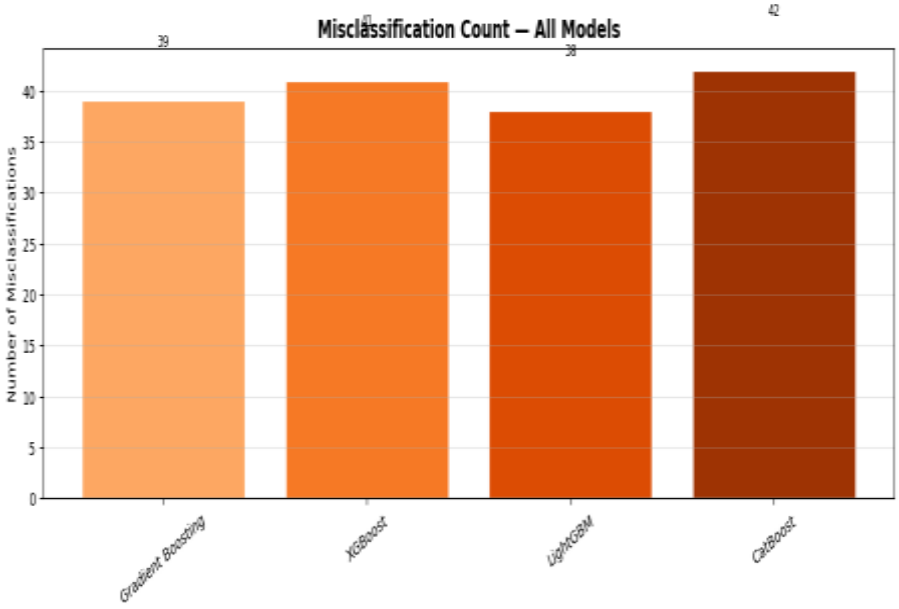

Fig. 15. Misclassification Count — Number of Incorrectly Classified Samples per Model.

VIII. DISCUSSION

The experimental results demonstrate that TF-IDF combined with gradient boosting ensemble classifiers is highly effective for automated fake news detection on the Kaggle benchmark. All four classifiers exceed 99.5% accuracy, with LightGBM marginally outperforming the others across most metrics due to its leaf-wise tree growth strategy, which minimizes loss more aggressively per boosting iteration. The near-identical performance across all models (F1 range: 0.0004) suggests that the TF-IDF feature representation is the dominant factor determining classification quality on this dataset.

The EDA reveals that fake and real news articles share a large vocabulary but differ in word frequency distributions and co-occurrence patterns. Bigrams such as "united states" and "white house" appear prominently in fake news, while "donald trump" and "hillary clinton" are top bigrams in real news — reflecting how each class discusses the same political figures but in different linguistic contexts. TF-IDF with ngram_range=(1,2) effectively captures these co-occurrence patterns.

The LIME analysis reveals that the model has learned interpretable, linguistically meaningful patterns. The strong influence of "Reuters" as a Fake predictor suggests that fake articles frequently cite Reuters (likely impersonation or misattribution), a pattern the model successfully identifies. Sensational and emotionally charged vocabulary ("Wackos," "UNREAL") are identified as strong Fake signals, consistent with research on misinformation writing styles.

Limitations of the present study include: (1) the dataset is limited to English-language news from a specific time period (pre-2018), limiting generalizability to multilingual or contemporary scenarios; (2) the pipeline does not incorporate metadata features such as author credibility, publication source reputation, or social propagation signals; (3) adversarial fake news crafted to mimic authentic journalistic style may evade TF-IDF-based detection; and (4) the calibration analysis reveals that all models produce miscalibrated probability outputs, which may limit their usefulness in downstream probabilistic decision systems without post-hoc calibration.

Future work will investigate: (a) transformer-based contextual representations (BERT, RoBERTa) as replacement for TF-IDF; (b) ensemble stacking combining gradient boosting with deep learning models; (c) multi-modal detection incorporating image and metadata features; (d) post-hoc probability calibration using Platt scaling; and (e) cross-lingual fake news detection.

IX. CONCLUSION

This paper presented a comprehensive comparative study of four gradient boosting ensemble classifiers — Gradient Boosting Classifier, XGBoost, LightGBM, and CatBoost — for automated fake news detection on the Kaggle Fake and Real News Dataset. The proposed pipeline combines a systematic multi-step text preprocessing procedure with TF-IDF vectorization using unigrams and bigrams. All four classifiers were evaluated on 8,980 test samples using six performance metrics.

LightGBM achieved the best overall performance with an Accuracy of 99.58%, F1-Score of 0.9959, and ROC-AUC of 0.9996, followed closely by Gradient Boosting, XGBoost, and CatBoost — all exceeding 99.5% accuracy. The near-equivalent performance across models confirms that TF-IDF features provide an effective and robust representation for this task. LIME explainability analysis demonstrated that the models rely on linguistically interpretable features, identifying journalistic attribution terms (e.g., "Reuters") and sensational vocabulary as key discriminative signals. Error analysis confirmed LightGBM as the most accurate classifier with only 38 misclassifications out of 8,980 test samples.

These findings establish TF-IDF-based gradient boosting as a strong, efficient, and interpretable baseline for misinformation detection, providing a solid foundation for more advanced contextual and multimodal approaches in future research.

X. REFERENCES

[1] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36, Jul. 2017.

[2] H. Ahmed, I. Traore, and S. Saad, "Detecting opinion spam and fake news using n-gram analysis and semantic similarity," in Proc. Int. Conf. Information Systems Security and Privacy (ICISSP), 2017, pp. 253–265.

[3] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," in Proc. IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kyiv, 2017, pp. 900–903.

[4] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[5] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems (NIPS), vol. 30, 2017.

[6] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in Advances in Neural Information Processing Systems (NeurIPS), vol. 31, 2018.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[8] C. Bisaillon, "Fake and Real News Dataset," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset