Machine Learning and Deep Learning for Detecting Fake News in a Low-Resource Language

Elton Tata, Jaumin Ajdarim, Nuhi Besimi
Faculty of Contemporary Sciences and Technologies, South East European University, Tetovo, Republic of North Macedonia

***Abstract*—Fake news detection has become a major problem in the digital age. This study presents an improved machine learning technique that achieves 91.99% accuracy in predicting fake news detection within Albanian textual datasets, demonstrating an improvement over existing baseline methodology. The implemented learning model uses 54 features that are specific to the Albanian language, such as red flags, credibility signals, punctuation patterns, and linguistic features. The model is tested on a balanced dataset of 3,994 news articles aggregated in Albanian from various sources. We compare it to several baselines, such as LSTM networks (80.35% accuracy) and BERT-augmented Naive Bayes classifiers (88.36% accuracy). In our Albanian dataset and experimental setting, XGBoost achieved 91.99% accuracy, indicating strong performance under the evaluated scenario.**

***Keywords*—Fake news; Albanian language; machine learning; NLP**

I. Introduction

The spread of lies and misinformation across the world, on digital platforms, poses a serious threat to democratic processes, sound decision-making, and public discourse. While much research has focused on developing sophisticated fake detection systems for high-resource languages such as English, Chinese, and Spanish by Kevin et al [1], Albanian, which is considered a low-resource language in the computational linguistics literature, has not yet been widely addressed. However, it lacks the extensive natural language processing resources, advanced pre-trained models, and annotated datasets that are available for major languages. In this context, the lack of resources presents interesting challenges for the development of efficient fake news detection systems and opportunities to investigate whether less expensive deep learning architectures can be used to achieve competitive performance using traditional shallow machine learning-based approaches developed with appropriate expertise in this field. From a computational perspective, Albanian has a number of features that make fake news detection both interesting and challenging. Special characters, such as "ë" (schwa) and "ç" (voiceless palatal affricate), are the foundation of the suggested language and help maintain a semantically accurate mapping. Due to the vocabulary of the Albanian language, there are many different types of hybrid words used in commerce, knowledge, and the performing arts [2]. The structure of the fixed definite article and the complex choice of verbs are two other notable patterns that stand in stark contrast to Romance and Germanic languages [3]. These linguistic features require customized features that our language model implementations take into account and capture the specific patterns of Albanian in both authentic journalism and fake news text.

A. Research Goals and Motivation

Three main observations of the state of the art in computational disinformation detection form the basis of the motivation behind this work. First, even in Albanian-speaking communities, the use of digital news and social media platforms for information consumption is more common, even in the absence of automatically filtered software to assess the credibility of their content. Second, due to structural differences in writing styles, languages, and cultural features, Albanian cannot be directly detected by standard fake news detection systems for most spoken languages.

Third, new research on transformer-based methods, such as multilingual BERT and other similar architectures [4]-[6], has shown that they can be used to detect fake news and do other NLP tasks across languages. But these methods usually need a lot of pretraining, a lot of computer power, and a long time to train, making them impractical for quick testing and deployment in research settings with few resources.

Our goal is to investigate the benefits of Albanian-specific features for deep learning techniques and compare them with traditional machine learning-based techniques (with or without language-specific enhancements) and the careless implementation of pre-trained models.

B. Contributions to Research

In the context of low-resource natural language processing and computational disinformation detection, this work offers three main contributions.

First, we create and validate a comprehensive set of 54 Albanian-specific linguistic features designed to capture structural and stylistic differences between fake and authentic Albanian news content, drawing on discourse-level cues inspired by Rhetorical Structure Theory (RST) [7] as well as carefully selected lexical, punctuation, and compositional patterns. While the latter prohibits the inclusion of features with low discriminatory power, such as sentence length, the former focuses on important aspects of the data to reduce overloading tendencies. Both the theory and conventions of Albanian news writing in the context of computational linguistics serve as the foundation for these features.

Second, using a set of methods, including hyperparameter search, forward selection from multiple model variants, and advanced text vectorization (with n-grams of characters), we

demonstrate a precisely calibrated Boost model with 91.99% accuracy.

Third, we demonstrate that, in the experimental context of this study and for the detection of Albanian-language fake news, a meticulously optimized machine learning methodology can attain superior accuracy and markedly reduced computational expenses in comparison to the assessed deep learning benchmarks.

The following sections comprise the rest of this document: Section II provides a literature survey of the earlier research related to this work. Section III presents the research methodology for the proposed study. In Section IV, the experimental results are presented. Section V show conclusion and future work.

II. Related Work

A. Content-Based Fake News Detection Approaches

Computational methods for detecting fake news and disinformation have evolved significantly over the past decade. Early approaches focused on content-based textual analysis and linguistic cues, drawing from techniques such as spam filtering and authorship identification [8]. Rubin et al. [9] identified stylistic and rhetorical patterns that differentiate deceptive content from legitimate journalism, showing that fake news often exhibits emotionally charged language, simplified grammatical structures, a lack of authoritative references, and sensational vocabulary [10].

More recent studies have expanded fake news detection beyond pure content analysis to incorporate source credibility and hybrid modelling strategies. Surveys by Shu et al. [11] and our previous work [12] consistently categorize detection methods into content-based, network-based, and hybrid approaches, while Guo et al. [13] further distinguish knowledge-based, style-based, and propagation-based strategies, building on propagation modelling techniques [14]. These studies consistently highlight content-based textual analysis as a core and indispensable component of effective fake news detection systems, motivating our emphasis on linguistic feature engineering.

B. Deep Learning Approaches and Low-Resource Constraints

Recent research has explored deep learning approaches, including recurrent and transformer-based models, for fake news detection. Such models are capable of automatically learning latent representations from raw text and have demonstrated strong performance in high-resource settings [15]. Transformer-based architectures, particularly BERT and its derivatives, have further advanced text classification performance and have been successfully adapted for fake news detection tasks [16],[17]. However, these approaches typically require large labelled datasets and substantial computational resources, which limit their practical applicability in low-resource research environments.

Low-resource languages face persistent challenges due to limited annotated data, scarce linguistic tools, and reduced representation in large pre-trained language models [18],[19]. Empirical studies further show that multilingual transformer models exhibit uneven performance across languages, with effectiveness strongly influenced by data availability and typological similarity [20]. These limitations highlight the need for alternative approaches that remain effective under data- and resource-constrained conditions.

C. Classical Machine Learning in Low-Resource Settings

A recurring question in fake news detection research concerns the conditions under which traditional machine learning methods can match or outperform deep learning approaches. Recent empirical studies indicate that well-designed classical models can achieve competitive performance on small- to medium-sized datasets when combined with effective feature engineering [21].

In particular, gradient-boosted decision tree models [22], including XGBoost, LightGBM, and CatBoost, have demonstrated strong performance and computational efficiency on structured and textual data. Essa et al. [23] report that LightGBM achieves high accuracy while maintaining computational efficiency, while other studies show that classical ensemble-based approaches can achieve competitive fake news detection performance with significantly lower computational cost than deep learning models. When paired with text vectorization techniques such as TF-IDF and higher-order n-grams, these models effectively capture discriminative linguistic patterns, making them especially suitable for domain-specific and low-resource scenarios. Overall, the reviewed literature provides strong justification for our approach, which emphasizes interpretable, Albanian-specific feature engineering and optimized classical machine learning models for effective fake news detection under limited data and computational resources.

III. Methodology

This section explains the extraction of features, the vectorization of text, and the architecture of our models (Fig. 2).

A. Data Collection and Sources

The dataset used in this study consists of 3,994 news articles in Albanian, collected from a wide range of Albanian-language media sources with different political orientations, geographical origins and journalistic standards [24].

There are 1,998 real news items and 1,996 fake news items in the dataset, which are almost perfectly balanced.

Every article that could be fake was manually reviewed, and only those that all commentators agreed on were kept in the final data.

The data includes the article title, the full text with original punctuation and paragraph structure, the date and time of publication, source information, and a binary class label.

B. Data Preprocessing and Quality Control

Text preprocessing steps were applied to reduce noise while preserving Albanian-specific linguistic characteristics relevant for downstream feature extraction. At the same time, this process explicitly preserves Albanian-specific characters, such as “ë” (schwa) and “ç”, which carry semantic value and are essential for the accurate representation of Albanian news texts (Fig. 1).

We address another Albanian-specific stopword removal, using a selected set of frequently occurring particles in texts with minimal discriminative value for classification, such as "dhe" (and), "e", "në" (in), "është" (is), "të" (to/of), "për" (for), "me" (me), "nga" (nga), etc., and their function words.

C. Data Partitioning Strategy

We use stratified random sampling to perform train-test splits and to ensure global class distribution in both subsets. 3,195 articles (80% of the data) are in the training set, with 1,598 articles that are real and an equal number that are fake, in which the class distribution remains approximately 50-50. The testing set includes 799 articles (20% of the total data), consisting of 400 real news items and 399 fake news items, once again maintaining class balance.

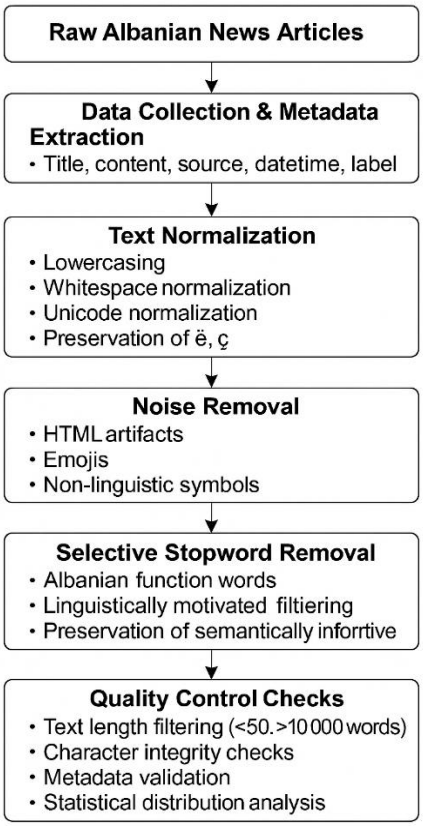

Fig. 1. Diagram for data processing and quality control.

D. Albanian-Specific Feature Engineering

We worked on 54 task-based features in five groups, taking different textual cues and metadata that we found empirically relevant for this active learning task - extracting different aspects of fabricated versus real news content.

E. Linguistic and Lexical Aspects of Albanian

The Albanian language has unique characteristics that enable the determination of genuine Albanian text versus text that may be translated from other languages or by non-native speakers [25]. The Albanian character ratio is the percentage of text that consists of two characters specific to Albanian, 'ë' and 'Ç', calculated as the number of these special characters divided by the total number of characters.

The binary variables control for the presence of Albanian characters (has_albanian_chars), which provides an admittedly simple but useful piece of information about the origin of the text. Vocabulary analysis features to check for the existence of sensational keywords that are typical of fake news in Albanian, such as "scandal" (scandal), "terror" (terror), "bomba" (bombë), "frikë" (fear), "shok" (shock), "alarm", "rrezik" (danger), "katastrofë" (catastrophe), "ekskluzive" (exclusive), "sekrete" (secret), "breaking" (borrowed English), "urgent" (borrowed English), "shikoni!" (look!) and "nuk do ta besoni!" (you won't believe it).

We derive both absolute counts (sensational word count) and ratios per document relative to document size (sensational ratio) to measure the strength of sensationalism, taking into account different lengths of articles.

The linguistic elements of opinion and defense cover question words and uncertainty devices such as "a" (whether), "vallë" (perhaps/I wonder), "ndoshta" (maybe), "mbase" (perhaps), "duket" (it appears), and "besoj" (I believe).

F. Characteristics of Punctuation Patterns

Patterns of punctuation [26] usage are strong indicators of Fake News vs. Professional News Sensationalism. Characteristics of exclamation marks consist of count, ratio of full document count to full document length, and binary flags for multiple exclamation marks (!! or !!!).

Question mark–based findings mirror the exclamation mark patterns. We analyze the number of question marks, their ratio relative to document length, and the detection of repeated question sequences (“??” or “???”). We track the use of ellipses, including in its sensationalist version, three consecutive dots (...). Contextual Citation and Citation Analysis.

Capitalization patterns are analyzed through the frequency and ratio of fully capitalized words, which tend to be more common in fake news than in professional journalism.

G. Statistical Textual Features

Word counts and sentence structures in the language model properties of the text are distinctive signals without being sensitive to the actual vocabulary [27]. We derive 15 statistical features to encode these features. Simple length measures are, e.g., total number of words, total number of characters, average word length (i.e., number of characters divided by number of words), and text length variation.

Sentence structure analysis applies regular expressions to determine sentence endings based on final punctuation (periods, exclamation marks, and question marks) and then calculates the number of sentences and their average word length, as well as their variation across the document. The unique word ratio represents the percentage of words confined to a single occurrence (type-mark) in the document, indicating diversity. The word repetition score, conversely, reflects the inverse metric (1 minus the type-mark ratio), which indicates how repetitive the content is.

H. Structural and Compositional Features

Aspects represent dependencies between content parts and logical structure [28]. We derive 8 features that control these

compositional properties. Headline analysis attributes that are analyzed independently of article titles include title length, title word count, and title punctuation (including exclamation and question marks). Fake news headlines tend to use question forms and their exclamatory counterparts to arouse curiosity or elicit an emotional response, while professional news headlines are more oriented toward so-called declarative patterns.

The title-content relationship models the relationship between headlines and news article bodies. Title_content_ratio calculates the ratio of title length to content, highlighting whether an article title is unusually long or short in relation to the information contained. Headline–content similarity features quantify the alignment between article titles and their corresponding bodies, capturing cases where titles diverge in length or content emphasis from the main text.

I. Time Features

When news timestamps are available, time features introduce an additional distinctive signal [16]. Five-time attributes have been identified from the publication_datetime metadata. The time of day takes the hour (values ranging from 0 to 23) when the article was published that day, which is expected to reflect different patterns in publication times. The day of the week (0="Monday", etc.) indicates the day of the week, allowing for weekly patterns in publication. The binary time indicators extracted are the is_weekend flag for Saturday/Sunday publications, capturing differences between weekend versus weekday publication patterns, and night by is_night if a publication was published between 22:00 and 05:00 local time when professional newsrooms are not active. This can reveal whether there are any patterns in the distribution of our data over time.

J. Construction of the Combined Feature Matrix

Each item is given 15,000 TF-IDF features using the vectorizer and 54 engineered_feats computed as described in the above section, resulting in a total data dimensionality of 15,054 features for each item. Scipy sparse matrix operations are used to join these feature sources into a single dataset by converting the 54 engineered features into dense arrays and stacking them horizontally with TF-IDF sparse matrices.

The gradient boosting model adapts the best feature weighting and interaction between both types of features, i.e., feature selection and synthesis, based on a unified training procedure.

K. Evaluation Methodology

We extract features for each test item in identical ways as during training and apply the trained TF-IDF vectorizer with a fixed vocabulary and IDF weights to transform the text and concatenate them to obtain 15,054-dimensional representations. We then use the models to produce probabilistic prediction results. The dimensionality was selected as a balance between representational richness and computational efficiency, as preliminary experiments indicated diminishing performance gains beyond this threshold.

Performance Metrics. Model performance is evaluated using standard classification metrics, including accuracy, precision per class, recall, and F1 score.

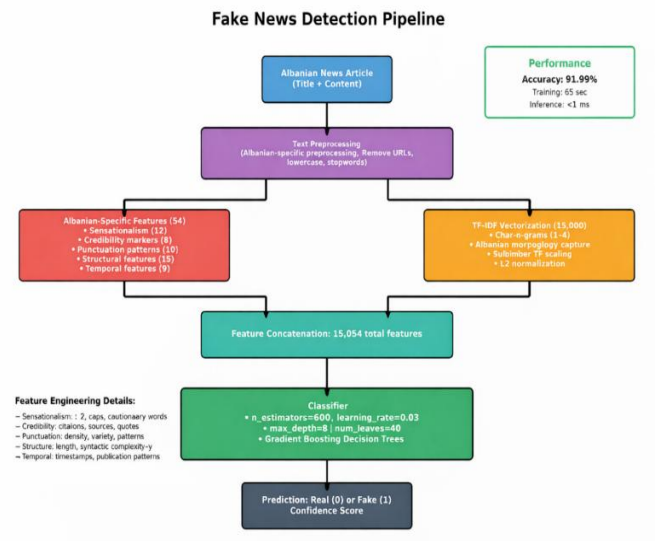

Fig. 2. Data processing, feature extraction, and classification pipeline used in this study.

IV. Experimental Results

The evaluation includes basic deep learning representational models, hybrid approaches, and gradient boosting models, to compare their performance under the same experimental conditions.

A. Summary of Evaluated Models

The evaluated models are grouped into four categories: basic deep learning models, hybrid models that combine neural representations with classical classifiers, gradient boosting models, and ensemble configurations. This categorization allows for a structured comparison of the models in terms of predictive performance and practical applicability.

B. LSTM Recurrent Neural Network and BERT-Enhanced Naive Bayes Hybrid

On the retained test set, the LSTM-based model achieved an accuracy of 80.35%, a significantly lower result than the best-performing gradient boosting models evaluated in this study. The comparatively lower performance of the LSTM model may be attributed to the moderate dataset size and the high dimensionality of the input representation. Deep sequential models typically benefit from substantially larger corpora or pretrained contextual embeddings. In low-resource settings, feature-engineered gradient boosting approaches may therefore exhibit greater stability and generalization capacity. With a training corpus of 3,195 items, the model seeks to learn word representations from scratch, which is insufficient for deep neural architectures that typically require significantly larger datasets to generalize well.

This hybrid model achieved an accuracy of 88.36%, a significant improvement of approximately 8 percentage points over the baseline LSTM model.

C. Logistic Regression Baseline

Despite its linear nature, logistic regression achieves competitive performance, achieving an accuracy of 89.36%, closely matching the result achieved by Naïve Bayes improved

by BERT. This result confirms that the combination of TF-IDF vectorisation and carefully designed Albanian language-specific features provides a very informative representation, even for simple linear classifiers.

D. Random Forest Baseline

The model achieved an accuracy of 88.49%, outperforming the baseline LSTM model (80.35%) and performing comparably to the BERT-enhanced Naïve Bayes model (88.36%).

The results indicate that tree-based ensemble methods combined with effective feature engineering can provide competitive performance without relying on deep neural architectures. The class-wise evaluation shows balanced behaviour, with true news accuracy of 88.91% and recall of 87.75%, and false news accuracy of 88.08% and recall of 89.22%, suggesting that there is no strong bias towards either class.

E. XGBoost Implementation

XGBoost is configured with such parameters for a fair comparison, including 500 boosting rounds, a 0.05 learning rate, a maximum tree depth limit of 7, a minimum child weight of 10, a subsample ratio of 0.8, and a column subsample ratio also set to 0.8. L1 regularization (alpha=0.1) and L2 regularization (lambda=1.0) are used to avoid overfitting on high-dimensional TF-IDF features by penalizing model complexity.

The XGBoost model achieved an accuracy of 91.99% on the full feature set. XGBoost model's prediction feature analysis indicates that it does well on articles with a clear vocabulary gap between fake and real news, can generalize across different-length articles by leveraging tree-based feature selection and captures non-linear interactions among predictive features with hierarchical splits (Fig. 3).

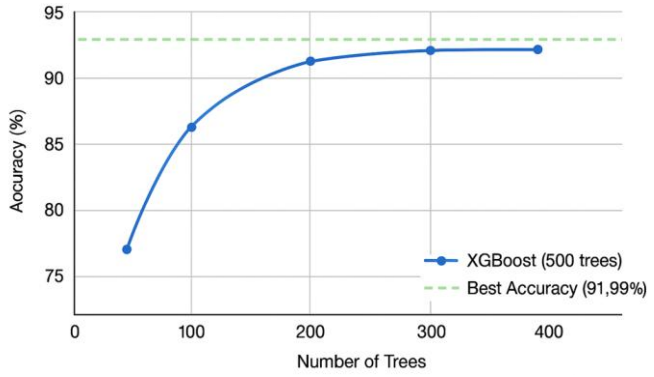

Fig. 3. Test accuracy of gradient boosting models.

F. Feature Importance and Interpretability Analysis

To further enhance the interpretability of the engineered feature set, feature importance analysis was conducted using XGBoost gain scores (Fig. 4). The results indicate that structural alignment cues, particularly the title_content_ratio, are among the most influential predictors. Content length and title-based attributes also rank highly, emphasizing that structural coherence plays a central role in distinguishing reliable journalism from misleading or clickbait-oriented content. Punctuation-related characteristics, including the frequency of exclamation marks and the presence of interrogative headlines, further contribute to classification decisions, supporting the hypothesis that exaggerated stylistic elements are associated with misleading articles. Lexical diversity measures such as word_repetition_score, unique word ratio, and word_length_variance also demonstrate notable predictive influence. Temporal attributes exhibit comparatively lower importance values, suggesting that they function as complementary contextual indicators rather than dominant classification drivers.

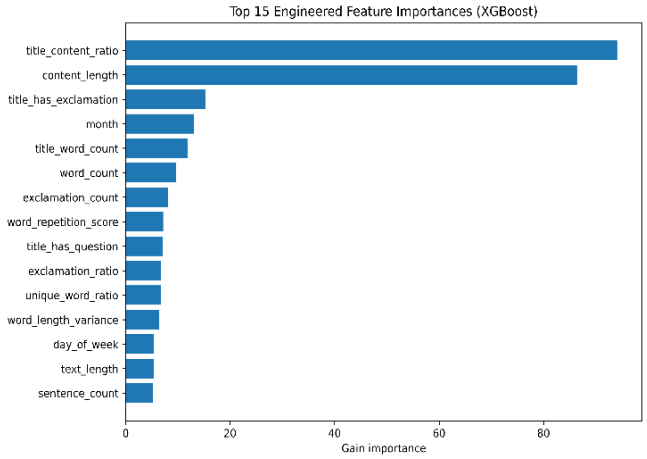

Fig. 4. Top 15 engineered feature importances based on XGBoost gain values.

G. Detailed Performance Evaluation

To provide a deeper evaluation of the classification behavior beyond the reported accuracy (91.99%), confusion matrix, and ROC–AUC analyses were conducted. These metrics offer additional insight into class-level performance and threshold sensitivity. The confusion matrix (Fig. 5) demonstrates a balanced distribution of classification outcomes across real and fake news categories. The model maintains comparable levels of false positives and false negatives, indicating the absence of systematic bias toward either class and confirming stable predictive behavior.

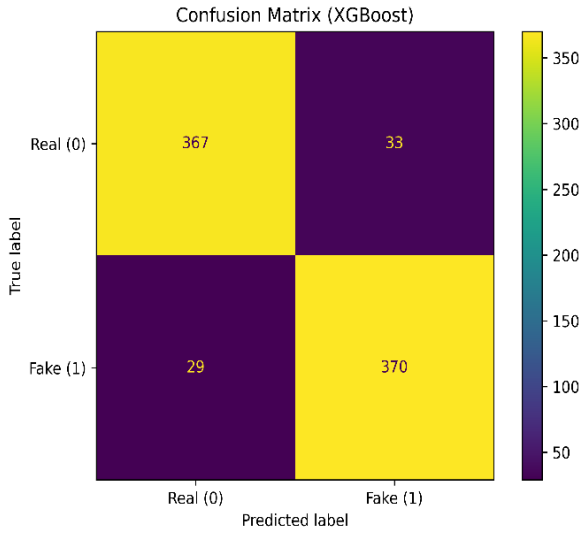

Fig. 5. Confusion matrix of the XGBoost classifier.

The Receiver Operating Characteristic (ROC) curve, Fig. 6, further illustrates strong separability between classes across varying decision thresholds. The model achieves an Area Under the Curve (AUC) score of 0.974, reflecting high discriminative capacity and robustness against threshold variation.

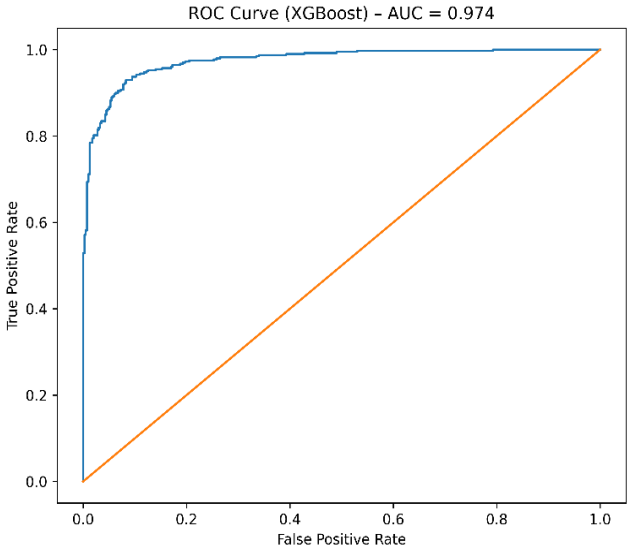

Fig. 6. ROC curve for the XGBoost classifier (AUC = 0.974).

H. Cross-Validation Stability Analysis

To ensure that the reported performance is not dependent on a single 80/20 train–test partition, a 5-fold stratified cross-validation procedure was conducted on the training data. The model achieved a mean accuracy of 91.08% with a standard deviation of 0.89%, indicating stable performance across folds. The low variance confirms that the classifier maintains consistent predictive behavior and does not rely on a specific data split. These findings support the robustness and generalization capacity of the proposed feature-engineered framework.

I. Error Analysis of Misclassified Instances

To further investigate model limitations, an analysis of misclassified instances was conducted based on the confusion matrix results. The model exhibits a balanced number of false positives and false negatives, indicating no systematic bias toward either class. False positives predominantly correspond to articles that display emotionally expressive or opinion-style language while remaining factually accurate. These cases suggest that stylistic intensity alone may sometimes resemble misleading patterns. On the other hand, false negatives usually include false information that is given in a neutral or formal tone and doesn't have any obvious sensational markers. Such cases highlight the challenge of detecting subtle or stylistically restrained misinformation.

Overall, the observed misclassifications suggest that ambiguity in stylistic signals remains the primary classification challenge, particularly when misleading content avoids exaggerated linguistic cues.

J. Comprehensive Model Comparison Table

Table I summarizes performance across all evaluated approaches, enabling direct comparison of accuracy, training time, model characteristics, and key strengths/weaknesses.

TABLE I. PERFORMANCE ACROSS ALL EVALUATED APPROACHES

Model	Accuracy	Training Time	Model Size	Key Strengths	Key Limitations
LSTM Baseline	*80.35%*	*60 min*	*2.4MB*	*Learns sequential patterns; no manual features*	*Slow training; overfits small data; requires large datasets*
Logistic Regression	*89.36%*	*1 sec*	*<1MB*	*Very fast; simple baseline; interpretable*	*Linear decision boundary; limited expressiveness*
BERT+NB Hybrid	*88.36%*	*2 min*	*850KB*	*Fast training; BERT tokenization*	*No contextual embeddings; independence assumption*
Random Forest	*88.49%*	*2 sec*	*15MB*	*Non-linear; handles interactions; interpretable*	*Outperformed by gradient boosting; larger model size*
XGBoost	*91.99%*	*35 sec*	*24MB*	*Robust; well-established*	*Best performance of the tested models.*

K. Comparative Analysis with Other Work

While this study focuses on fake news detection in Albanian, it is important to contextualize the obtained results in relation to existing studies conducted in other languages and datasets. A direct comparison with previous work allows a deeper understanding of whether the achieved performance is consistent with established empirical findings or represents an isolated result.

Therefore, a comparative analysis is conducted with representative fake news detection studies that use similar experimental environments, including textual datasets of comparable size and widely used machine learning and deep learning approaches. The comparison focuses mainly on accuracy, which is the most widely reported evaluation metric in the relevant literature.

As shown in Table II, previous studies on fake news detection report accuracy values that generally range between approximately 88% and 95%, depending on the data characteristics, feature representations, and learning models used. Classical machine learning approaches, especially ensemble-based models such as Random Forest and XGBoost, consistently demonstrate strong performance on medium-scale textual data.

TABLE II. COMPREHENSIVE MODEL PERFORMANCE COMPARISON

Study	Dataset Size	Method	Accuracy
This work (Albanian Fake News Detection)	*3,994*	*LR*	*89.36%*
		XGBoost	*91.99%*
		LSTM	*80.35%*
		BERT with Naive Bayes	*88.36%*
Alameri & Mohd [28]	*~5,000–7,000*	*SVM*	*88.20%*
		Naive Bayes	*84.60%*
		LSTM	*94.21%*
Abdulrahman & Baykara [29]	*7,796*	*Naive Bayes*	*90.10%*
		Random Forest	*95.66%*
		CNN	*97–98%*
Ahmad & Lokeshkumar [30]	*20,000*	*SVM*	*92.00%*
		Naive Bayes	*90.00%*
		Neural Network	*93.00%*
Alghamdi et al. [31]	*LIAR (~12k)*	*LR*	*88.50%*
		SVM	*89.00%*
		Naive Bayes	*87.00%*
		XGBoost	*92.10%*
		LSTM	*91.00%*
		BERT	*93.40%*
Albahr & Albahar [32]	*LIAR (~12k)*	*Naive Bayes*	*89.00%*
		Random Forest	*87.60%*
Sharma et al. [33]	*~6,000*	*LR*	*88.90%*
		Naive Bayes	*86.20%*
		Random Forest	*90.80%*
Kumar et al. [34]	*1,356*	*LSTM*	*88.00%*
		CNN + BiLSTM + Attention	*88.78%*

Deep learning models perform competitively when ample training data is present; however, their efficacy may fluctuate in contexts with constrained datasets. In this context, this study achieves an accuracy of 91.99% using TF-IDF features combined with XGBoost, which aligns well with the results reported in the literature for comparable experimental conditions.

These findings show that the proposed approach follows established empirical trends observed in different languages and datasets, confirming that gradient-boosted machine learning models remain an effective and reliable choice for fake news detection, especially in languages with limited data resources.

L. Ethical Considerations

Automated fake news detection systems may carry risks of misclassification that could affect journalistic credibility and public trust. Therefore, such systems should be deployed as decision-support tools rather than fully autonomous labeling mechanisms. Transparency, interpretability, and periodic human oversight are essential to ensure responsible use, particularly in low-resource linguistic contexts such as Albanian.

V. CONCLUSION

In this work, we showcased an optimized machine learning model for Albanian fake news detection, which achieved 91.99% accuracy by combining optimal Albanian feature engineering and XGBoost. This work shows that generic machine learning methods can achieve comparable performance to deep-learning solutions through integrating domain-specific knowledge and seeking systematic optimization.

We introduced and validated 54 Albanian specific optimization, linguistic features capturing sensationalism indicators, credibility markers, punctuation patterns, structural information and temporal features. The optimized XGBoost achieves an accuracy of 91.99%, significantly higher than other models.

REFERENCES

[1] Martínez-Gallego, K., Álvarez-Ortiz, A. M., & Arias-Londoño, J. D. (2021). Fake news detection in spanish using deep learning techniques. *arXiv preprint arXiv:2110.06461*.

[2] Krasniqi, K. (2019). The relation between language and culture (Case study Albanian Language). *Linguistics and Literature studies*, *7*(2), 71-74.

[3] Biberauer, T., & Roberts, I. (2008). Subjects, tense and verb-movement in Germanic and Romance. *Cambridge occasional papers in linguistics*, *3*, 24-43.

[4] Anirudh, K., Srikanth, M., & Shahina, A. (2023, December). Multilingual fake news detection in low-resource languages: A comparative study using BERT and GPT-3.5. In *International Conference on Speech and Language Technologies for Low-resource Languages* (pp. 387-397). Cham: Springer Nature Switzerland.

[5] De, A., Bandyopadhyay, D., Gain, B., & Ekbal, A. (2021). A transformer-based approach to multilingual fake news detection in low-resource languages. *Transactions on Asian and Low-Resource Language Information Processing*, *21*(1), 1-20.

[6] Bala, A., & Krishnamurthy, P. (2023, September). Abhipaw@ dravidianlangtech: Fake news detection in dravidian languages using multilingual bert. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages* (pp. 235-238).

[7] Vargas, F., Jonas, D. A., Rabinovich, Z., Benevenuto, F., & Pardo, T. (2022, June). Rhetorical structure approach for online deception detection: A survey. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 5906-5915).

[8] Martins, R., Almeida, J. J., Henriques, P., & Novais, P. (2021). A sentiment analysis approach to improve authorship identification. *Expert Systems*, *38*(5), e12469.

[9] Rubin, V. L., Conroy, N., Chen, Y., & Cornwell, S. (2016, June). Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection* (pp. 7-17).

[10] Wick-Pedro, G., de Sales Santos, R. L., & Vale, O. (2024, November). Linguistic and emotional dynamics in satirical vs. real news: a psycholinguistic analysis. In *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)* (pp. 386-392). SBC

[11] Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, *53*(5), 1-40.

[12] Tata, E., Ajdari, J., & Besimi, N. (2023, May). Fake News Detection: A Comprehensive Survey. In *2023 46th MIPRO ICT and Electronics Convention (MIPRO)* (pp. 309-314). IEEE.

[13] Guo, Z., Schlichtkrull, M., & Vlachos, A. (2022). A survey on automated fact-checking. *Transactions of the association for computational linguistics*, *10*, 178-206.

[14] Tufchi, S., Yadav, A., & Ahmed, T. (2023). A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and

opportunities. *International Journal of Multimedia Information Retrieval*, *12*(2), 28

[15] Bahad, P., Saxena, P., & Kamal, R. (2019). Fake news detection using bi-directional LSTM-recurrent neural network. *Procedia Computer Science*, *165*, 74-82.Shishah, W. (2021). Fake news detection using BERT model with joint learning. *Arabian Journal for Science and Engineering*, *46*(9), 9115-9127.

[16] Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia tools and applications*, *80*(8), 11765-11788.

[17] Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *arXiv preprint arXiv:2004.09095*

[18] Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2021, June). A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2545-2568).

[19] Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT?. *arXiv preprint arXiv:1906.01502*.

[20] Alameri, S. A., & Mohd, M. (2021, January). Comparison of fake news detection using machine learning and deep learning techniques. In *2021 3rd international cyber resilience conference (CRC)* (pp. 1-6). IEEE

[21] Lin, J., Tremblay-Taylor, G., Mou, G., You, D., & Lee, K. (2019, December). Detecting fake news articles. In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 3021-3025). IEEE

[22] Essa, E., Omar, K., & Alqahtani, A. (2023). Fake news detection based on a hybrid BERT and LightGBM models. *Complex & Intelligent Systems*, *9*(6), 6581-6592.

[23] Canhasi, E., Shijaku, R., & Berisha, E. (2022). Albanian fake news detection. *Transactions on Asian and low-resource language information processing*, *21*(5), 1-24

[24] Kabashi, B. (2018). A lexicon of Albanian for natural language processing. *Lexicographica*, *34*(1), 239-248.

[25] Sousa-Silva, R. (2022). Fighting the fake: A forensic linguistic analysis to fake news detection. *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique*, *35*(6), 2409-2433.

[26] Samadi, M., Mousavian, M., & Momtazi, S. (2021). Deep contextualized text representation and learning for fake news detection. *Information processing & management*, *58*(6), 102723.

[27] Uppal, A., Sachdeva, V., & Sharma, S. (2020, January). Fake news detection using discourse segment structure analysis. In *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 751-756). IEEE.

[28] Alameri, S. A., & Mohd, M. (2021, January). Comparison of fake news detection using machine learning and deep learning techniques. In *2021 3rd international cyber resilience conference (CRC)* (pp. 1-6). IEEE.

[29] Abdulrahman, A., & Baykara, M. (2020, December). Fake news detection using machine learning and deep learning algorithms. In *2020 international conference on advanced science and engineering (ICOASE)* (pp. 18-23). IEEE.

[30] Ahmad, F., & Lokeshkumar, R. (2019). A comparison of machine learning algorithms in fake news detection. *International Journal on Emerging Technologies*, *10*(4), 177-183.

[31] Alghamdi, J., Lin, Y., & Luo, S. (2022). A comparative study of machine learning and deep learning techniques for fake news detection. *Information*, *13*(12), 576.

[32] Albahr, A., & Albahar, M. (2020). An empirical comparison of fake news detection using different machine learning algorithms. *International Journal of Advanced Computer Science and Applications*, *11*(9).

[33] Sharma, U., Saran, S., & Patil, S. M. (2020). Fake news detection using machine learning algorithms. *International Journal of creative research thoughts (IJCRT)*, *8*(6), 509-518.

[34] Kumar, S., Asthana, R., Upadhyay, S., Upreti, N., & Akbar, M. (2020). Fake news detection using deep learning models: A novel approach. *Transactions on Emerging Telecommunications Technologies*, *31*(2), e3767.