Automated Detection of Albanian Gegë and Toskë Dialects Through Classical and Deep-Learning Classifiers

Eliot Bytyçi
Department of Mathematics, University of Prishtina, Kosovo
eliot.bytyci@uni-pr.edu

Enda Alidema
Department of Mathematics, University of Prishtina, Kosovo
enda.alidema@student.uni-pr.edu

Jetmir Gjoni
Department of Mathematics, University of Prishtina, Kosovo
jetmir.gjoni@student.uni-pr.edu

Trime Ismajli
Department of Mathematics, University of Prishtina, Kosovo
trime.ismajli@student.uni-pr.edu

Abstract: *This study presents a framework for automatically distinguishing between the two major Albanian dialects, Gegë and Toskë, using a hybrid of classic classifiers and deep-learning classifiers. Through the course of the study, we have collected a balanced corpus of dialectal samples from social media posts and online texts. After preprocessing of the dataset, initially three classical algorithms were evaluated: Extreme Gradient Boosting (XGBoost), due to its robust overfitting control and non-linear performance; Support Vector Machine (SVM), due to margin-based classification in high-dimensional text spaces; and a Recurrent Neural Network with Long Short-Term Memory (RNN-LSTM), to capture sequential dependencies in dialectal language. Moreover, after transforming raw text into numerical features, we have compared embeddings from four pretrained multilingual models: Bidirectional Encoder Representations from Transformers (BERT)-base-multilingual-cased, Nomic-embed-text-v2-moe, Jina-embeddings-v3, and Cross-lingual Language Model-Robustly Optimized BERT Pretraining Approach (XLM-RoBERTa-large). Each embedding-classifier pairing was assessed on accuracy and F1-score. Final results, showed that XGBoost from classical and XLM-RoBERTa-large from deep learning, achieved the highest accuracy. Nevertheless, other algorithms showed promising results such as RNN-LSTM and BERT, which excelled at modeling sequential patterns. Moreover, SVM offered a compelling balance of performance and efficiency, demonstrating the practical value of using classical algorithms instead of usage of advanced embeddings with targeted classifiers for dialect identification.*

Keywords: *Albanian language dialects, natural language processing, machine learning, deep learning.*

Received August 08, 2025; accepted December 28, 2025
https://doi.org/10.34028/iajit/23/5/1

1. Introduction

The Albanian language, as the language of Albanians, constitutes a distinct unique branch of the Indo-European family, representing the sole surviving member of the Albanoid subfamily with origins in ancient Paleo-Balkan tongues [9]. Its phylogenetic position and long continuity render Albanian a pivotal subject in comparative Indo-European studies. Today, it is spoken by an estimate of over 7 million native speakers and holds official status in Albania and Kosovo, co-official status in North Macedonia and Montenegro, and recognition in minority communities across Serbia, Italy, Croatia, Romania, and Greece.

Albanian language exhibits two principal regional dialects-Gegë (northern part) and Toskë (southern part) with the Shkumbin River forming the traditional division between them [4]. Although standard Albanian is codified on Toskë basis, its usage on digital platforms such as social media, news sites, and forums remains pervasively influenced by both dialects [16]. Dialectal heterogeneity poses significant challenges for Natural Language Processing (NLP) because dialectal forms often diverge from prescriptive norms.

The explosion of user-generated digital text, such as Tweets, Facebook posts, blog entries, and online article offers vast, naturally occurring corpora that intimately reflect authors' dialectal backgrounds rather than the standardized language form. While this data abundance presents opportunities for sociolinguistic modeling, the historic scarcity of large, annotated corpora, pre-trained models, and comprehensive toolkits for Albanian has impeded progress in complex NLP tasks such as dialect classification. Overcoming these constraints demands methods capable of learning from limited data and capturing subtle lexical, morphological, and syntactic markers distinguishing Gegë from Toskë.

The challenges of NLP for "low resource" languages, unlike English or German, Albanian has historically lacked large, annotated datasets (corpora), pre-trained models, and comprehensive NLP toolkits. This scarcity of resources makes complex tasks like dialect classification even more challenging. It requires innovative methods that can learn from limited data and capture the subtle lexical, morphological, and syntactic differences between the dialects.

Initial research in Albanian NLP concentrated on foundational resources morphological analyzers, Part-of-Speech (POS) taggers, and n-gram classifiers, often evaluated on modest, homogeneous datasets. Recent efforts have extended these foundations: a geographically informed multi-dialectal corpus derived from Twitter demonstrated the feasibility of large-scale dialect annotation and revealed dialect-specific features previously undocumented.

This paper significantly advances prior work by presenting a large-scale, balanced corpus of Gegë and Toskë texts compiled from social media, news outlets, and existing datasets. We systematically benchmark a suite of classification models from classical baselines (Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM)) to deep sequential Recurrent Neural Network with Long Short-Term Memory (RNN-LSTM) and transformer-based embeddings paired with state-of-the-art multilingual models (Bidirectional Encoder Representations from Transformers (BERT)-multilingual, Nomic MoE, Jina AI, Cross-lingual Language Model Roberta Robustly Optimized BERT Pretraining Approach (XLM-RoBERTa)). Finally, we conduct an in-depth feature and error analysis to illuminate the key linguistic markers that computationally differentiate the two dialects.

Specifically, our contributions are threefold:

1. We compile and present a new, a larger-scale corpus of Gegë and Toskë texts drawn from a diverse range of sources, including social media and news articles, additionally also from previous papers.
2. We implement and benchmark a wide array of classification models, from established baselines to state-of-the-art deep learning architectures.
3. We provide a detailed feature analysis and error analysis that offers new insights into the key linguistic markers that differentiate the two dialects computationally

The remainder of the paper is organized as follows. Section 2 reviews related work in dialect identification and low-resource language processing of the Albanian language. Section 3 details our data collection, annotation processes, and modeling methodology. Section 4 presents experimental results, comparative performance evaluations, and discussion. Section 5 concludes with key insights and outlines avenues for future research.

2. Related Work

The rapidly expanding body of Albanian-language NLP research still lags work on higher resource languages, yet recent studies show steady progress in adapting different techniques to the constraints of a low-resource setting. Below, initially we surveyed two application areas: hate-speech detection and news/sentiment classification, and then followed by an overview of dialect identification in other languages that informs our own Albanian dialect study.

2.1. Hate-Speech and Offensive-Language Detection

There are three interesting papers that have used classification, for the task of hate speech detection. Zenuni *et al*. [18] presented the first attempt to create a hate speech corpus and classifier for Albanian language, collecting 4,886 comments from two Facebook pages and annotating them as hate or non-hate using two annotators. The authors implemented a SVM classifier using RTextTools in R, splitting their dataset into 4,000 training samples and 886 testing samples to perform binary classification of hate speech content. Their SVM-based approach achieved a precision of 0.61, recall of 0.57, and F1-score of 0.58, representing pioneering work in Albanian hate speech detection despite the relatively modest performance results.

Fetahi *et al*. [6], tackled limited research that has been conducted on hate speech detection in the Albanian language, taking initial steps in developing an automated hate speech classifier for Albanian texts using the SVM algorithm, achieving an F1-score of 0.58. While, Nurce *et al*. [13] investigate both classical machine learning and deep learning models using their established Albanian dataset for offensive speech detection, with their BERT transformer model experiments achieving an accuracy of 0.77.

Ajvazi and Hardmeier [1] present a novel dataset for offensive language detection in Albanian social media, collecting 3,000 user-generated comments from Kosovo news platforms on Facebook and YouTube and annotating them using the hierarchical Offensive Language Identification Dataset (OLID) scheme. Their fine-tuned multilingual BERT model achieved a macro-F1-score of 0.86 for Albanian offensive language classification, significantly outperforming previous Albanian datasets that reported F1-scores of 0.58-0.80. The authors' transfer learning experiments revealed that combining Albanian and Danish training data improved performance on Danish classification tasks but showed diminished results for Albanian, suggesting language-specific characteristics in offensive language patterns.

Moreover, in the age of Large Language Models (LLMs), there is also a possibility of AI generated text and thus we need research that helps distinguishing that text from real one. Hazim and Ata [7] propose machine learning-based NLP framework that effectively distinguishes AI-generated from human-written academic texts, achieving very high accuracy.

2.2. News Categorization and Sentiment Analysis

Plaku *et al*. [14] present a machine learning framework for automated classification of news article titles in Albanian, addressing the challenges posed by limited

text corpora and the complexity of the Albanian language. A dataset of 9600 news titles across six categories is introduced, and various machine learning algorithms, particularly Recurrent Neural Networks (RNNs), are evaluated for their effectiveness in classifying these titles, achieving accuracy of 90%. The study demonstrates that deep learning methods outperform traditional classifiers (logistic regression, random forest, etc.,) in accurately categorizing news articles in low-resource languages like Albanian.

Nuci *et al*. [12] present EduSenti, a corpus comprising 1,163 Albanian and 624 English reviews of educational instructors' performance annotated for sentiment, emotion, and educational topic, representing the first work to apply sentiment analysis to students' feedback in the education domain for Albanian. Nuci *et al*.'s [12] experiment with fine-tuning several language models including pretrained Albanian embeddings from XLM-RoBERTa checkpoints, achieving F1-scores of 71.9 for Albanian and 73.8 for English on sentiment classification tasks. Their work demonstrates the feasibility of sentiment analysis in low-resource languages like Albanian, though they note that Albanian-specific pretrained models underperformed compared to multilingual XLM-RoBERTa, suggesting the need for improved pretraining strategies and larger datasets for this language.

Kadriu *et al*. [10] compare two approaches for Albanian text classification using news articles from five Albanian portals: a bag-of-words model with nine scikit-learn classifiers and fastText for word analogies based on character n-grams. Their evaluation on 1,430 articles across seven categories (latest, economy, sport, showbiz, technology, culture, world) reveals that the perceptron classifier achieves the highest accuracy of 94% using the bag-of-words approach, significantly outperforming fastText's 86% accuracy on single-label classification tasks. However, fastText demonstrates superior performance for multi-label text classification, achieving 79% accuracy on a larger South East European-SETimes corpus with 160 categories, while traditional classifiers struggled with multi-label scenarios, achieving only 45% accuracy at best.

Yet Kastrati and Biba [11] present the first comprehensive survey of NLP for the Albanian language, reviewing foundational tools POS taggers, morphological analyzers, stemmers, parsers, and annotated corpora) and core applications (named-entity recognition, sentiment and emotion analysis, hate speech detection, summarization, text classification, and question answering) across rule-based, machine learning, and deep learning paradigms. They report that early rule and dictionary-based methods achieved high precision (POS tagging up to 97%) while recent deep learning approaches (e.g., Bidirectional Long Short-Term Memory (BiLSTM) with attention for sentiment analysis) deliver improved flexibility but are constrained by limited annotated data, yielding F1-scores from 58% (hate speech detection) to 92% (sentiment analysis). The authors identify critical gaps, such as the need for larger, publicly available corpora, language-specific tokenizers, and domain-tailored pretrained models and call for coordinated data collection and tool release to advance Albanian NLP research.

2.3. Dialect Identification Beyond Albanian Language

Dialect classification has matured in other languages, offering transferable insights. Most of the paper we checked where in Arabic, where for example a country-level tasks were experimented through SVM with dictionary-based features and pointwise mutual information values [3] achieving an average F1-score of almost 19%. Priban and Taylor [15] experimented with Arabic dialects, in order to detect the dialect and predict the country of the tweet origin. Alsuwaylimi [2] used BiLSTM to classify binary between Arabic dialects, with accuracy of around 87%. This substantial improvement highlighting that recurrent architectures could learn complex sequential dependencies that hand-engineered features captured incompletely and thus undergoing transformative shift toward dialect-specific transformer models.

In other studies, especially in Chinese language, a study by Hu *et al*. [8] discussed variations between mainland and Taiwan variations by adopting base classifiers with ensemble methods, to achieve results of around 0.9 F1-score. Xu *et al*. [17] used general features like character-level n-gram, and many new word-level features, including PMI-based and word alignment-based features, to achieve better results. This ensemble success demonstrates that no single feature set or architecture captures all relevant dialectal distinctions; rather, combining lexical, orthographic, and learned semantic representations provides more robust discrimination

Another paper by Ciobanu *et al*. [5], this time in German dialect identification, based on SVM classifiers, reached the results of 62% of F1-score. But, more recently the researchers are moving also towards neural approaches, thus demonstrating a critical insight: for languages with limited digitized corpora, transfer learning from large multilingual models substantially improves performance over task specific training.

The reviewed literature highlights the growing use of classification techniques for several languages, with emphasize in using transformer technology, nowadays. Since, most existing NLP research for Albanian has focused on tasks like hate speech detection and sentiment analysis, little work has addressed dialect classification. As a contrast, as seen from above review, dialect identification has been a well-studied problem in other languages. Thus, that was our motivation as well, to use classical models of machine learning but also the

ones based on transformers, to see if they can effectively capture subtle syntactic and lexical differences across dialects, offering a methodological foundation for low-resource languages like Albanian. And we believe that patterns mentioned above tell us that transformers have advantages, even though in some cases for low resource languages, simple models sometimes perform very good. Thus, more motivation to use both. Then data scarcity and class imbalance as constraints, can be addressed through data augmentation and transfer learning strategies. Another addition to our motivation, was to use well established methodologies mentioned from other languages.

Our study thus deliberately employs both classical machine learning approaches and transformer-based methods, mirroring methodological diversity evident across languages. This dual approach establishes whether classical methods with n-gram features remain competitive for Gegë-Toskë, evaluates transformer models to understand multilingual pretraining effectiveness for Albanian specific distinctions, and explores ensemble combinations, learning from Chinese and other experiences that hybrid architectures often outperform single models. The persistent performance gap between Gegë and Toskë mirrors documented challenges across other low resource dialect pairs, underscoring necessity of the data augmentation and staged enhancement strategies discussed.

3. Description of Data, their Collection and Models Used

This section presents a comprehensive and scientifically rigorous account of the dataset preparation, preprocessing protocols, and machine learning model selection developed for the computational analysis of the Gegë and Toskë dialects of the Albanian language. The methodological framework encompasses systematic data sourcing, rigorous data validation, advanced feature engineering, and contemporary embedding strategies, all with the aim of accurately capturing and modeling nuanced linguistic divergences between these dialects.

3.1. Dataset Construction and Validation

The primary corpus originated from multiple authentic and representative sources, including 1,660 Gegë dialect words extracted from the Albanian literary work Lahuta e Malësisë, as well as digitized texts sourced from online libraries and publicly accessible social media platforms (Facebook, Twitter). Each word from the Gegë dialect was meticulously paired with its corresponding Toskë (standard Albanian) equivalent to ensure precise dialectal contrast.

This study acknowledges the limited size of the corpus, which is common in low-resource dialect research but presents inherent constraints for deep learning approaches. With approximately 1,660-word pairs, this dataset serves as a preliminary foundation for Gegë-Toskë dialect analysis rather than a comprehensive representation of the full linguistic variation across both dialects. The modest dataset size limits the generalizability of findings and may result in reduced model robustness, particularly for complex downstream NLP tasks. As such, this work should be interpreted as exploratory research on a limited dialect corpus, establishing baselines and methodological foundations for future investigation.

While the class distribution is balanced in terms of word pair counts, a persistent performance gap exists between the Gegë and Toskë representations. This asymmetry likely reflects inherent biases stemming from the imbalanced linguistic resources available for each dialect. Gegë, as dialect, has fewer digitized resources and less representation in contemporary written corpora compared to Toskë (standard Albanian). This data scarcity results in less diverse contextual examples for Gegë words, potentially limiting model learning and generalization for the minority dialect.

To validate the integrity and linguistic fidelity of the word pairs, a multi-tiered review process was implemented:

1. Initial cross-referencing with trusted resources such as Fjalorthi
2. Consultation with expert Albanian linguists for expert verification
3. Systematic comparison with publicly available dialectal dictionaries and reputable repositories.

Emphasis was placed on incorporating commonly used lexical items to maximize practical relevance and reliability of the dataset for dialectological analysis. Future research should consider multiple complementary strategies to address data scarcity and dialectal imbalance. These include:

1. Data augmentation techniques such as back-translation to artificially expand the Gegë subset.
2. Synthetic data generation methods (e.g., Synthetic Minority Over-sampling TEchnique (SMOTE) or generative models) to create linguistically plausible additional examples of underrepresented dialectal forms.
3. Expanded corpus collection through systematic annotation of social media content and collaboration with Gegë language communities to gather authentic contemporary language use.

Additionally, employing cost-sensitive learning approaches or staged learning strategies that gradually balance class representation during training may help mitigate the performance disparity between dialects. Community engaged corpus development, in collaboration with native speakers and local language experts, represents a particularly promising direction for

sustainable growth of Albanian dialect resources. For our Gegë-Toskë corpus with around 1,660-words, thus given the limited dataset size, the most appropriate split we thought would be 70% for training, with 15% for validation and 15% for testing.

3.2. Preprocessing and Data Cleaning

A robust, multi-stage preprocessing pipeline was established to enhance data quality and analytical rigor. Initial normalization involved the removal of punctuation and conversion of all text to lowercase, followed by the elimination of superfluous characters (e.g., hashtags, emojis) frequently encountered in social media texts. Advanced lexical tokenization procedures were applied, segmenting textual entries into individual tokens while preserving dialect-specific contractions and morphological features critical for linguistic analysis. Superior quality control measures were enforced, including automated and manual duplicate detection, anomaly filtering to exclude mixed-dialect entries, and outlier removal. This approach ensured standardization of the dataset while maintaining sensitivity to features salient for dialect discrimination.

This study employs a purposeful two-stage research design that establishes a foundational baseline using raw dialectal word forms before introducing advanced morphological preprocessing techniques, reflecting established best practices in NLP for new language pairs and dialect resources. The primary rationale for this foundational approach is threefold: to establish interpretable baselines raw form embeddings provide a transparent, reproducible starting point against which morphological preprocessing can be systematically evaluated; to minimize confounding variables when developing resources for a low-resource language pair like Gegë-Toskë, introducing multiple simultaneous preprocessing techniques makes it difficult to isolate which factors contribute to performance improvements or degradation; and to preserve information preservation while morphological preprocessing (lemmatization, stemming) inevitably removes surface-level information that may encode dialectal markers, orthographic conventions, or fine-grained linguistic distinctions critical to dialect differentiation.

Moreover, this study is explicitly framed as exploratory research on a limited corpus, establishing methodological foundations for Gegë-Toskë computational dialectology, and the modest dataset size justifies adoption of a simpler baseline approach that prioritizes interpretability and foundational understanding over optimized performance. Future research phases will systematically introduce morphological processing through a planned progression: implementing Byte-Pair Encoding (BPE) tokenization, applying lemmatization and morphological tagging, investigating morphology-aware embedding architectures.

3.3. Machine Learning Model Design and Selection

Given the dataset's scale (small to moderate-sized) and the inherent challenges of textual classification, three well-established machine learning paradigms were selected to collectively address overfitting, feature complexity, and sequential dependencies:

- XGBoost: chosen for its advanced gradient boosting architecture, recognized for robustness against overfitting and proficiency in modeling complex, non-linear relationships frequently observed in textual data.
- SVM: well-suited to high-dimensional feature spaces, the margin-maximizing characteristics of the SVM make it a strong candidate for dialect classification tasks in linguistically diverse datasets.
- RNN-LSTM: utilized for its capacity to model sequential context and long-range dependencies, enabling the capture of syntactic and morphological patterns distinctive to dialectal variance.

3.4. Embedding and Vectorization Approaches

To facilitate meaningful comparison between the Gegë and Toskë lexical items, four advanced pretrained multilingual embedding models were systematically evaluated for their efficacy in capturing semantic and lexical nuances:

- BERT-base-multilingual-cased (Google): employs a contextualized transformer-based architecture, allowing for the generation of dynamic sentence representations sensitive to dialect-specific context.
- Nomic-embed-text-v2-moe: leverages a mixture-of-experts paradigm to enhance semantic expressiveness and adaptability across languages.
- Jina-embeddings-v3 (Jina AI): targets cross-lingual similarity modeling, thus offering potential advantages for comparative dialectal studies.
- XLM-RoBERTa-large (Facebook AI): a robust cross-lingual language model trained on over 100 languages, facilitating broad linguistic generalization.

Each embedding approach was quantitatively and qualitatively assessed against the Gegë and Toskë dialect dataset, focusing on three principal evaluation criteria: the capacity to support nuanced Albanian dialect features, the ability to generate meaningful vector representations, and the measurement of lexical and semantic divergence through cosine similarity metrics. This multifaceted methodological paradigm provided a robust foundation for the computational analysis of dialectal variation within Albanian. Moreover, while fine-tuning transformer models represents a common practice in NLP, this study strategically employed pre-trained transformers (multilingual BERT, XLM-RoBERTa) as fixed feature

extractors rather than pursuing full fine-tuning. This methodological choice reflects both theoretical and practical considerations particularly relevant to low-resource dialect pairs.

4. Experimental Results

This research presents a comprehensive evaluation of machine learning and deep learning methodologies for Albanian dialect classification, employing rigorous statistical assessment protocols and advanced algorithmic approaches. Multiple classification paradigms were systematically investigated, including traditional ensemble methods such as XGBoost, kernel-based approaches like SVM, and RNN architectures, specifically RNN.

The experimental validation protocol employed standard machine learning evaluation metrics, with accuracy serving as the primary performance indicator alongside precision, recall, and F1-score assessments. Initial empirical results demonstrated that RNN architectures achieved superior classification performance, attaining an accuracy of 79.88%, representing a statistically significant improvement over traditional machine learning approaches, presented in Table 1 and visualized in Figure 1.

Table 1. Machine learning algorithms.

Classical algorithm	Results		
	Accuracy	F1-score (Gegë)	F1-score (Toskë)
XGBoost	73.12%	0.63	0.79
SVM	73.12%	0.65	0.78
RNN	79.88%	0.77	0.82

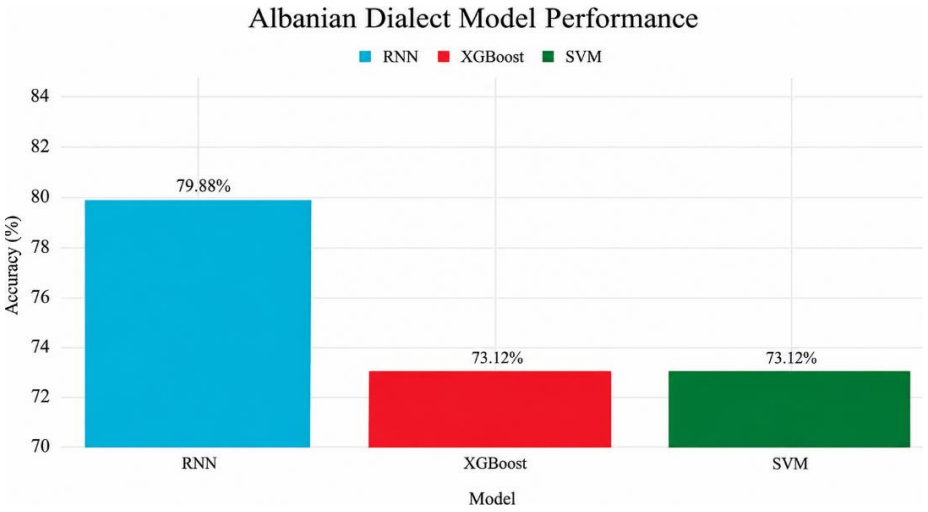

Figure 1. Albanian dialect model performance.

Both XGBoost and SVM methodologies yielded comparable performance levels at 73.12% accuracy, indicating a performance differential of 6.76 percentage points favoring the deep learning approach.

However, comprehensive error analysis and confusion matrix evaluation revealed systematic classification challenges, particularly in the accurate identification of the Gegë dialect. This performance degradation can be attributed to the significantly higher morphological complexity inherent in the Gegë dialect system compared to its Toskë counterpart. The Gegë dialect exhibits extensive inflectional variability, including complex vowel systems with nasal vowels, morphophonological alternations, consonant cluster simplifications, and distinctive verbal conjugation patterns that create substantial feature space dimensionality challenges for computational models.

The observed classification difficulties with morphologically rich dialectal variants align with established findings in computational linguistics regarding the inherent challenges posed by morphologically complex languages in NLP tasks. Morphologically rich languages typically demonstrate increased lexical sparsity due to extensive inflectional paradigms, resulting in a high proportion of low-frequency word forms that impede effective statistical learning. This phenomenon, known as the morphological complexity paradox, creates particular challenges for both traditional feature-based approaches and neural network architectures in achieving robust generalization across dialectal variations.

The superior performance of RNN models can be attributed to their sequential processing capabilities, which enable the capture of long-range dependencies and contextual patterns characteristic of natural language sequences. Unlike traditional machine learning approaches that rely on static feature representations, RNN architectures process text as sequential data, allowing for the modeling of temporal dependencies and morphosyntactic relationships that are particularly relevant for dialect classification tasks. However, the persistent challenges encountered with the Gegë dialect suggest that even advanced sequential models struggle with the extreme morphological variability present in this dialectal system, indicating the need for specialized preprocessing techniques, morphologically-aware tokenization strategies, or hybrid architectural approaches that can better

accommodate morphological richness.

To enhance generalization capacity and extract more sophisticated linguistic representations, we systematically evaluated four state-of-the-art pretrained multilingual embedding architectures: BERT-base-multilingual-cased, Nomic-embed-text-v2-moe, Jina-embeddings-v3, and XLM-RoBERTa-large. Each textual sequence underwent vectorization through these respective embedding models, followed by classification using a lightweight logistic regression classifier, employing the embeddings as dense feature representations for discriminative learning.

The experimental evaluation protocol employed comprehensive performance metrics including overall classification accuracy, precision, recall, and dialect-specific F1-scores to assess both balanced classification performance and class-wise discrimination effectiveness. The comparative analysis revealed substantial performance variations across embedding architectures, as detailed in the following empirical findings, and presented in Figure 2.

Figure 2. Performance comparison of multilingual embedding models on Albanian dialect classification.

Jina-embeddings-v3 demonstrated superior performance across all evaluation dimensions, achieving 82% overall accuracy with well-balanced dialectal discrimination capabilities (F1-score: 0.79 for Gegë, 0.84 for Toskë). This multilingual embedding model's effectiveness can be attributed to its task-specific Low-Rank Adaptation (LoRA) architecture and matryoshka representation learning framework, which enables flexible embedding dimensionality while maintaining semantic fidelity across diverse linguistic contexts.

Both XLM-RoBERTa-large and BERT-base-multilingual-cased exhibited competitive performance levels, each attaining 79% classification accuracy with comparable F1-scores (XLM-RoBERTa: 0.77 Gegë, 0.82 Toskë; BERT: 0.78 Gegë, 0.82 Toskë). The robust performance of XLM-RoBERTa can be attributed to its extensive pretraining on over 100 languages, providing enhanced cross-lingual representation capabilities particularly beneficial for under-resourced languages such as Albanian. However, despite BERT's demonstration of high intra-dialectal semantic coherence (average cosine similarity>0.84), the model exhibited insufficient fine-grained discriminative capacity between closely related dialectal variants, limiting its classification efficacy.

Nomic-embed-text-v2-moe achieved moderate classification performance with 76% accuracy and corresponding F1-scores of 0.74 for Gegë and 0.77 for Toskë. The mixture-of-experts architecture, while theoretically advantageous for capturing diverse linguistic patterns, demonstrated minimal semantic separation between dialectal embeddings (average cosine similarity: 0.598, embedding distance: 0.067). This limited discriminative capacity suggests that the model's architectural complexity may not effectively leverage dialectal variation patterns within the Albanian linguistic space.

The systematic performance advantage for Toskë dialect classification across all embedding models reflects the inherent linguistic and data-driven biases present in the evaluation framework. Toskë serves as the foundation for standard Albanian and maintains greater representation in curated multilingual datasets, resulting in more robust learned representations and consequently higher classification performance. Conversely, the Gegë dialect classification challenges stem from its substantial morphological variability, informal register usage, and limited standardization, creating feature space complexities that impact recall performance, particularly in traditional and mixture-of-experts models.

These empirical findings demonstrate that deep multilingual embedding architectures, particularly Jina-embeddings-v3 and XLM-RoBERTa-large, which are specifically optimized for semantic clustering and

cross-lingual representation learning, provide significant performance improvements for dialectal classification tasks. The results underscore that model effectiveness depends not only on embedding quality and semantic representation capacity but also on the classification algorithm's ability to exploit subtle morphosyntactic and lexical variations inherent in Albanian dialectal grammar. Furthermore, the consistent superior performance on Toskë dialect classification highlights the critical importance of balanced training data representation and the need for specialized preprocessing techniques to address morphological complexity in dialectologically diverse linguistic datasets.

5. Conclusions, Limitations and Future Work

This research presents a comprehensive computational approach to Albanian dialect classification, systematically evaluating both traditional machine learning methods and contemporary deep learning architectures across multiple embedding paradigms. The investigation demonstrates significant advances in automated dialectal discrimination while revealing fundamental challenges inherent to morphologically complex language varieties.

The empirical findings establish Jina-embeddings-v3 as the superior approach for Albanian dialect classification, achieving 82% accuracy with balanced performance across both Gegë and Toskë dialects.

This represents a substantial improvement over traditional machine learning methods, with RNN architectures achieving 79.88% accuracy compared to 73.12% for XGBoost and SVM approaches.

The research contributes to the broader understanding of computational dialectology by demonstrating that embedding quality significantly impacts classification performance, particularly for under-resourced language varieties. The systematic evaluation of four state-of-the-art multilingual embedding models provides valuable insights into the relative effectiveness of different architectural approaches for dialectal discrimination tasks. Furthermore, the identification of morphological complexity as a primary constraint for automated dialect classification advances theoretical understanding of the computational challenges posed by inflectionally rich linguistic systems.

Despite these contributions, several significant limitations constrain the generalizability and practical applicability of the findings. Dataset composition bias represents a fundamental challenge, with Toskë dialect achieving consistently higher classification accuracy due to its standardization and greater representation in training corpora. This imbalance reflects broader issues in computational linguistics research for under-resourced languages, where standard varieties receive disproportionate attention compared to dialectal variants.

Morphological preprocessing limitations constitute another critical constraint, as current approaches lack specialized techniques for handling the extensive inflectional variability characteristic of the Gegë dialect. The absence of morphologically-aware tokenization and feature extraction methods limits model effectiveness in capturing the full range of dialectal variation present in spontaneous speech and informal text.

Scale constraints further limit the research impact, as the dataset size (1,660-word pairs) represents a relatively small corpus for deep learning approaches, potentially limiting model generalization across diverse dialectal contexts. This constraint is compounded by the lack of specialized Albanian linguistic resources, including morphological analyzers, POS taggers, and syntactic parsers optimized for dialectal variation.

Evaluation methodology limitations include the focus on lexical-level classification rather than discourse-level or pragmatic features that may provide additional discriminative power for dialectal identification. Additionally, the binary classification framework (Gegë vs. Toskë) does not capture the full spectrum of regional variation within each major dialect group.

Several promising research directions emerge from these findings that could substantially advance computational dialectology for Albanian and similar morphologically complex languages. Morphologically-aware preprocessing represents the most critical area for future development, requiring the creation of specialized tokenization algorithms, morphological analyzers, and feature extraction techniques tailored to Albanian dialectal variation. This includes developing robust methods for handling code-switching, morphophonological alternations, and dialect-specific inflectional patterns.

Balanced dataset creation and augmentation constitutes another priority area, involving systematic collection of Gegë dialect materials from diverse sources and the development of data augmentation techniques that preserve linguistic authenticity while expanding training corpus size. This should include the creation of balanced corpora representing different registers, domains, and regional sub-varieties within each major dialect group.

Hybrid architectural approaches offer significant potential for improving dialectal classification performance by combining the strengths of different model types. Future research should investigate architectures that integrate morphological awareness, sequential processing capabilities, and multilingual embedding representations in unified frameworks. This includes exploring attention mechanisms specifically designed for dialectal variation and multi-task learning approaches that jointly optimize for multiple linguistic tasks.

Cross-dialectal transfer learning represents an underexplored area with substantial potential for improving performance on low-resource dialectal varieties. Future work should investigate techniques for leveraging linguistic knowledge from related dialects and languages to improve classification performance, particularly for informal and code-switched content.

Integration with broader Albanian NLP infrastructure would enhance the practical utility of dialectal classification systems by incorporating them into comprehensive language processing pipelines. This includes developing dialect-aware machine translation systems, sentiment analysis tools, and information retrieval applications that can adapt to dialectal input.

Evaluation framework expansion should encompass more sophisticated metrics that capture not only classification accuracy but also linguistic plausibility, cultural sensitivity, and practical utility in real-world applications. This includes developing evaluation protocols that assess model performance across different text types, domains, and user populations.

The advancement of Albanian computational dialectology requires sustained interdisciplinary collaboration between computational linguists, Albanian language specialists, and community stakeholders to ensure that technological developments serve the needs of Albanian-speaking communities while preserving linguistic diversity and cultural authenticity. Future research should prioritize community-centered approaches that involve native speakers in dataset creation, evaluation, and application development to ensure cultural and linguistic appropriateness.

Acknowledgment

This work was supported by the joint grant of MEST of Kosovo and HEI25 project.

References

[1] Ajvazi A. and Hardmeier C., "A Dataset of Offensive Language in Kosovo Social Media," *in Proceedings of the 13th Language Resources and Evaluation Conference*, Marseille, pp. 1860-1869, 2022. https://aclanthology.org/2022.lrec-1.198/

[2] Alsuwaylimi A., "Arabic Dialect Identification in Social Media: A Hybrid Model with Transformer Models and BiLSTM," *Heliyon*, vol. 10, no. 17, pp. 1-18, 2024. https://doi.org/10.1016/j.heliyon.2024.e36280

[3] Althobaiti M., "Country-Level Arabic Dialect Identification Using Small Datasets with Integrated Machine Learning Techniques and Deep Learning Models," *in Proceedings of the 6th Arabic Natural Language Processing Workshop*, Kyiv, pp. 265-270, 2021. https://aclanthology.org/2021.wanlp-1.30/

[4] Cerpja A. and Cepani A., "Albanian Dialect Classifications," *Dialectologia: Revista Electrònica*, vol. 11, pp. 51-87, 2023. https://www.raco.cat/index.php/Dialectologia/article/view/427733

[5] Ciobanu A., Malmasi S., and Dinu L., "German Dialect Identification Using Classifier Ensembles," *in Proceedings of the 5th Workshop on NLP for Similar Languages, Varieties and Dialects*, New Mexico, pp. 288-294, 2018. https://aclanthology.org/W18-3933/

[6] Fetahi E., Hamiti M., Susuri A., Ajdari J., and Zenuni X., "AI-based Hate Speech Detection in Albanian Social Media: New Dataset and Mobile Web Application Integration," *International Journal of Interactive Mobile Technologies*, vol. 18, no. 24, pp. 190-208, 2024. https://online-journals.org/index.php/i-jim/article/view/50851

[7] Hazim L. and Ata O., "Bridging the Gap: Ensemble Learning-Based NLP Framework for AI-Generated Text Identification in Academia," *The International Arab Journal of Information Technology*, vol. 22, no. 6, pp. 1055-1069, 2025. https://doi.org/10.34028/iajit/22/6/2

[8] Hu H., Li W., Zhou H., Tian Z., and et al., "Ensemble Methods to Distinguish Mainland and Taiwan Chinese," *in Proceedings of the 6th Workshop on NLP for Similar Languages, Varieties and Dialects*, Michigan, pp. 165-171, 2019. https://aclanthology.org/W19-1417/

[9] Jashari A., "On the Structure of Adjectives' Superlative Degree Formation by Means of Analytical Tools: A Contrastive Analysis of Albanian and some other Indo-European Languages," *Knowledge-International Journal*, vol. 69, no. 5, pp. 847-852, 2025. https://ojs.ikm.mk/index.php/kij/article/view/7308

[10] Kadriu A., Abazi L., and Abazi H., "Albanian Text Classification: Bag of Words Model and Word Analogies," *Business Systems Research*, vol. 10, no. 1, pp. 74-87, 2019. DOI: 10.2478/bsrj-2019-0006

[11] Kastrati M. and Biba M., "Natural Language Processing for Albanian: A State-of-the-Art Survey," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 6, pp. 6432-6439, 2022. http://doi.org/10.11591/ijece.v12i6.pp6432-6439

[12] Nuci K., Landes P., and Di Eugenio B., "RoBERTa Low Resource Fine Tuning for Sentiment Analysis in Albanian," *in Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, Torino, pp. 14146-14151, 2024. https://aclanthology.org/2024.lrec-main.1233/

[13] Nurce E., Keci J., and Derczynski L., "Detecting Abusive Albanian," *arXiv Preprint*, vol. arXiv:2107.13592v3, pp. 1-10, 2021. https://arxiv.org/abs/2107.13592

[14] Plaku E., Jahaj K., Cela A., and Civici N., "A Machine Learning Framework for Automated News Article Title Classification in Albanian," *in Proceedings of the International Conference on Innovations in Intelligent Systems and Applications*, Craiova, pp. 1-6, 2024. https://ieeexplore.ieee.org/document/10683815

[15] Priban P. and Taylor S., "ZCU-NLP at MADAR 2019: Recognizing Arabic Dialects," *in Proceedings of the 4th Arabic Natural Language Processing Workshop*, Florence, pp. 208-213, 2019. https://aclanthology.org/W19-4623/

[16] Toska M., Nivre J., and Zeman D., "Universal Dependencies for Albanian," *in Proceedings of the 4th Workshop on Universal Dependencies*, Barcelona, pp. 178-188, 2020. https://aclanthology.org/2020.udw-1.20/

[17] Xu F., Wang M., and Li M., "Sentence-Level Dialects Identification in the Greater China Region," *International Journal on Natural Language Computing*, vol. 5, no. 6, pp. 9-20, 2016. DOI: 10.5121/ijnlc.2016.5602

[18] Zenuni X., Ajdari J., Ismaili F., and Raufi B., "Automatic Hate Speech Detection in Online Contents Using Latent Semantic Analysis," *Press Academia Procedia*, vol. 5, no. 1, pp. 368-371, 2017. https://dergipark.org.tr/en/pub/pap/article/371817

Eliot Bytyçi is an Associate Professor at the University of Prishtina and a Researcher in Computer Science, with interests in Data Mining, Semantic Web, and AI/ML. He has contributed to multiple national and international research projects and has published widely in ICT-related fields.

Enda Alidema is a DevOps Engineer based in Pristina, with a background in Mathematics and Computer Science from the University of Prishtina. Her professional work focuses on DevOps Practice and Software Infrastructure. She holds Master of Science in Computer Science.

Jetmir Gjoni serves in the IT sector and has held leadership and technical roles in public-sector Technology Administration. His background includes Systems Administration and IT Management Experience, along with enrollment in a Master's degree in Computer Science.

Trime Ismajli is an AI/ML Engineer and Mathematician based in Pristina. She has a strong foundation in mathematics and applied machine learning. She holds Master of Science in Computer Science.