Classification of Albanian Social Media Posts into Toskë and Gegë dialects

Trime Ismajli[1,†], Jetmir Gjoni[1,†], Enda Alidema[1,†] and Eliot Bytyçi[1,*,†]

[1] *University of Prishtina, Avenue Mother Teresa, No-5, 10000, Prishtinë, Republic of Kosova*

Abstract

The Albanian language is characterized by two major dialectal variants, Toskë and Gegë, which show distinct linguistic features in vocabulary, pronunciation, and grammar. The rise of social media has increased the volume of user-generated content in both dialects, presenting a challenge for natural language processing (NLP) due to informal writing styles, code-switching, and the lack of structured datasets. This study aims to classify Albanian social media posts into Toskë and Gegë dialects using machine learning techniques, including XGBoost, Support Vector Machines (SVM), and Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM). As a secondary contribution we curated a dataset of 1,660 manually annotated words and sentences, collected from diverse sources including literature and social media, and pre-processed to ensure consistency. Using machine learning algorithms, we evaluated the performance of these models in dialect classification. Our findings reveal that Recurrent Neural Networks (RNNs) achieved an accuracy of 79.88%, outperforming both XGBoost and Support Vector Machines (SVMs). The study highlights the importance of addressing data shortage and linguistic variability in dialect classification, offering a foundational resource for future research in Albanian Natural Language Processing (NLP). This work paves the way for practical applications in machine translation, dialect-aware chatbots, and language preservation efforts, while also highlighting the need for expanded datasets and refined models to improve classification accuracy.

Keywords

Natural language processing, Albanian dialects, Machine learning algorithms

1. Introduction

The Albanian language is known for its rich dialectal diversity, primarily characterized by two major dialect variants: Toskë and Gegë [1]. These dialects exhibit distinct differences in vocabulary, pronunciation, and grammatical structures, reflecting the deep cultural and historical heritage of Albanian language. The author of [2] points out that the incredible variety of dialects in the Albanian language stems from its rich history and cultural exchanges. Meanwhile, [3] argues that the unique characteristics of the Toskë and Gegë dialects have emerged from centuries of linguistic growth and regional influences. Standard Albanian is based on the Toskë dialect, whereas Gegë is mostly spoken in northern Albania, Kosovo, Montenegro, and North Macedonia [4].

This linguistic heterogeneity is extremely challenging for Natural Language Processing (NLP), especially for tasks such as classifying social media posts that are informal. The absence of any structured annotated datasets for Toskë and Gegë dialects has made the development of meaningful classification models challenging [5]. Furthermore, the informal nature of social media posts also makes it more difficult for automatic dialect classification based on inconsistent spelling, use of slang,

[1] *6th International Conference Recent Trends and Applications in Computer Science and Information Technology RTA-CSIT, May 22-24, 2025, Tirana, Albania*

[*] Corresponding author.

[†] These authors contributed equally.

trime.ismajli@student.uni-pr.edu (T. Ismajli); jetmir.gjoni@student.uni-pr.edu (J. Gjoni); enda.alidema@student.uni-pr.edu (E. Alidema); eliot.bytyci@uni-pr.edu (E. Bytyçi)

0009-0008-6869-2062 (T. Ismajli); 0009-0008-6550-0387 (J. Gjoni); 0009-0006-2343-9034 (E. Alidema); 0000-0001-7273-9929 (E. Bytyci)

abbreviations, use of emojis, and code-switching [6]. Other challenges include the overlap in lexical items across dialects and the problem of unbalanced datasets.

Due to social media usage on Facebook, X (formerly Twitter), and Instagram, there is an increasing amount of user-generated content in both Albanian dialects. Dialect classification with high accuracy can greatly benefit NLP applications like machine translation, dialect-aware chatbot creation, and language preservation [7].

The primary objective of this study is to address these challenges by developing a robust model capable of classifying Albanian social media posts into Toskë and Gegë dialects. To achieve this, the research seeks to answer key questions, such as:

- How can a high-quality dataset be curated to support the classification of these dialectal variants?

This study aims to bridge these gaps by not only classifying social media posts into Toskë and Gegë dialects using machine learning models such as XGBoost, Support Vector Machines (SVM), and Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) layers but also contributing to the creation of a small, annotated dataset.

Thus, the contributions of this research are twofold. First, it provides a classification model capable of distinguishing between Toskë and Gegë dialects in social media posts, offering a valuable tool for linguistic analysis and NLP applications. Second, it introduces a curated dataset that addresses the scarcity of resources for these Albanian dialects, paving the way for further research and development in this area.

The remainder of this paper is organized as follows: Section 2 reviews related work on dialect classification using machine learning, as well as social media text analysis. Section 3 describes the dataset and methodology employed in this study. Section 4 presents the results of the research and discusses their implications. Finally, Section 5 concludes the paper and suggests potential directions for future research, emphasizing the importance of expanding resources and refining models for Albanian dialect classification.

2. Related work

The classification of dialects has become a major problem area in Natural Language Processing (NLP) and is attributed to language diversity and lack of dialect resources. Studies have shown that machine learning techniques can be helpful to address this problem in many languages. For example, Hamilton Neural Networks (HNNs) enhanced speech identification of Arabic dialects by incorporating temporal and spectral speech features. Similarly, some ensemble learning methods, particularly those using Linear Support Vector Machines, were successful in classifying the Hawrami dialect, proving again that model selection and feature selection are critical [8][9][7].

The main challenge in automation of dialect classification is the lack of annotated corpora, especially for minor dialects such as Binjhal, which inhibits the effort to build necessary NLP tools [8][9][10]. Incorporating social media is even more challenging because the level of noise introduced due to informal writing, switching between languages, word contractions, and misspellings is far greater than in other domains. Nevertheless, social media is a major source of dialectal information. Research on Arabic dialect classification has demonstrated that machine learning can be taught to work well with social media content, but extensive feature engineering, coupled with deep preprocessing, is a prerequisite [6][8][10].

One of the recent advances in Albanian NLP is creating a multi-dialectal Twitter corpus for Albanian, Kosovo, and North Macedonian dialects. When machine learning models were trained on the dataset mentioned above, they were extremely accurate in determining dialects, addressing the issue of data scarcity and demonstrating the potential of social media as a linguistic resource [11][12][10]. Additionally, Albanian question classification research using the TREC dataset confirmed that deep learning models were better than traditional methods, suggesting the applicability of these models to dialect classification tasks as well [12][13][11].

Complementary research in other low-resource settings provides additional information. Parmar and Bhavsar [13][14][12] proposed a deep ensemble structure that combined CNN, BiLSTM, and GRU as the Indo-Aryan dialect features. Using data augmentation techniques, their model was effective even with limited training data—a strategy which could be transferred to similar Albanian tasks. In addition, Zoubi and El-Beltagy [14][15][13] introduced a hybrid transformer-based model for Arabic dialect identification on social media. Their model integrated BERT embeddings with CNN and BiLSTM layers, which successfully tackled the issues of noisy and code-mixed text. More generally, transformer-based models (such as BERT and its multilingual relatives) have proven extremely successful for dialect identification, as their contextual word representations capture long-distance dependencies without any need for feature engineering special.

Despite these advances, gaps remain. Many studies continue to concentrate on high-resource dialects while leaving others underrepresented [9][10][8]. The unavailability of free, standardized, and curated social media datasets also limits generalizability towards dialect classification models. Additionally, the evolving nature of spontaneous online language demands more adaptive and noise-resilient modeling practices.

This study attempts to resolve the stated issues via the Toskë-Gegë dialect classification of Albanian social media content. Machine learning algorithms like XGBoost, SVM, and RNN with LST, are utilized on a hand-curated manually tagged dataset. Our vision is to contribute towards current research on dialect classification, particularly for low-resource settings, and to offer blueprints for subsequent NLP applications on the Albanian language.

3. Dataset and methodology

A significant challenge in Albanian NLP research is the scarcity of labeled datasets for dialect classification. To address this issue, this study contributes to the creation of a small but valuable annotated dataset containing words and sentences labeled as either Toskë or Gegë [15]. The dataset consists of:

- 1,660 words and sentences, collected from diverse sources such as literature (e.g., Lahuta e Malcisë), social media posts (Facebook, Instagram, Twitter) from January 2025 to March 2025 using the hashtags #Gegë and #Shqip, and publicly available online repositories.
- Manual annotation conducted by linguistic experts and cross-verified with online dictionaries such as Fjalorthi[2] and the Academy of Sciences Dictionary[3].
- Pre-processing steps, including text cleaning, lowercasing, removal of punctuation and non-alphabetic characters, and correction of spelling inconsistencies.

This dataset, publicly available on GitHub, serves as a foundational resource for future research in Albanian dialect processing, enabling further development of NLP applications such as automated translation systems and dialect-aware chatbots. The dataset is presented in CSV format.

By addressing both dialect classification and data scarcity challenges, this study not only advances computational linguistics research for the Albanian language but also paves the way for practical applications in machine translation, social media analytics, and digital communication.

After the manual annotation, we have gathered words in a column, with another column added to describe the word dialect, a needed step for classification. Then we cleaned the text for punctuation, symbols and other non-alphabetic characters, while also checking for any spelling typo.

[2] https://fjalorthi.com/

[3] http://m.fjalori.shkenca.org/

Later on, we tested if any empty or null submissions had appeared by accident. After that, since each of the Gegë items has a Toskë equivalent, we made sure that the dataset is perfectly balanced so that no dialect might be favored.

The algorithms chosen for the classification were:

- XGBoost: is considered more robust to overfitting and able to capture non-linear relationships (that exist when dealing with text), our dataset can be considered small to medium in which case XGBoost performs better [16].
- SVM: data has many features and SVM performs well with many features, is robust to overfitting and works well with sparse data.
- RNN: As a classification model already in Keras library which are designed to handle sequential data, have the Long Short-Term Memory (LSTM) layer suitable for text data to capture long-range dependencies in text.

We implemented our models on Google Colab and locally in PyCharm and uploaded them to the GitHub repository [15].

4. Experimental results

Below we are presenting some of the findings, starting with Table 1, showing that XGBoost and SVM are performing similarly and the best performance for the above mentioned dataset has RNN.

Table 1. Accuracy of all the algorithms

Algorithms	Accuracy
XGBoost	73.12%
SVM	73.12%
RNN	79.88%

Moreover, below we present the confusion matrix for XGBoost in Figure 1, then the confusion matrix of SVM in Figure 2 and the confusion matrix of RNN in Figure 3.

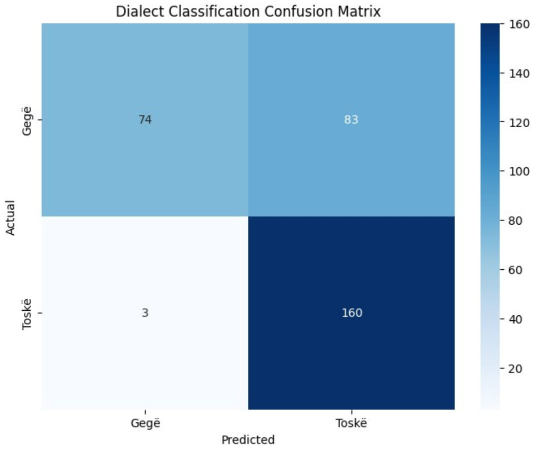

Figure 1. Confusion matrix for XGBoost

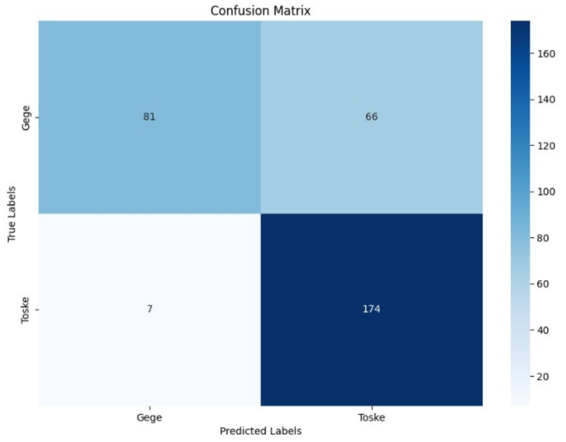

Figure 2. Confusion matrix for SVM

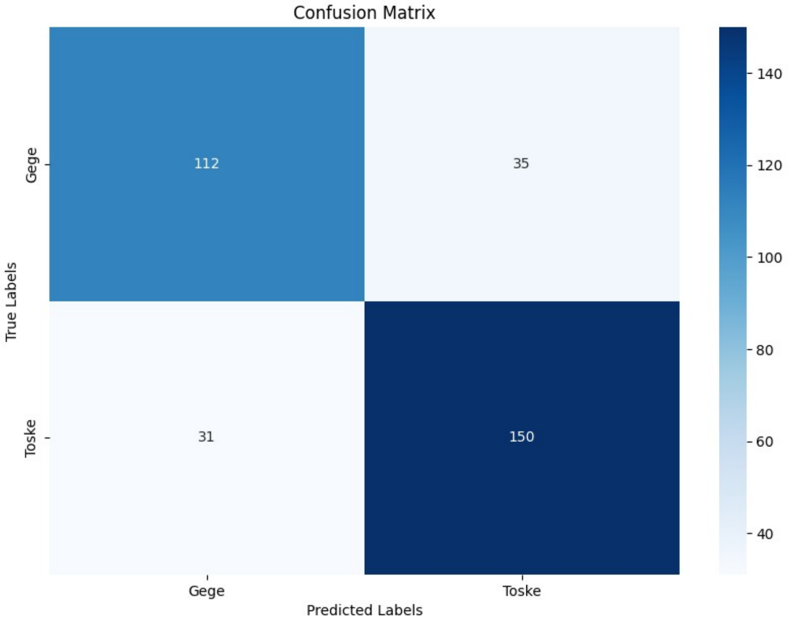

Figure 3. Confusion matrix for RNN

The classification report for the XGBoost model provides a detailed evaluation of its performance across two classes, labeled as Gege and Toske (or in figure described as 0 and 1). For class Gege, the model achieves a high precision of 0.96, indicating that when it predicts this class, it is correct 96% of the time. However, the recall is relatively low at 0.47, meaning it only identifies 47% of the actual instances of class Gege. This results in an F1-score of 0.63, which balances precision and recall. For class Toske, the model has a lower precision of 0.66 but a high recall of 0.98, showing it effectively captures 98% of the actual instances of this class, leading to a higher F1-score of 0.79. The overall accuracy of the model is 0.73, meaning it correctly classifies 73% of all instances. The macro and weighted averages for precision, recall, and F1-score are both 0.81, 0.73, and 0.71, respectively, indicating consistent performance across both classes when considering class imbalance, as presented in Table 2.

Table 2. Classification report for XGBoost

Class	Precision	Recall	F1 score	Support
0	0.96	0.47	0.63	157
1	0.66	0.98	0.79	163

While the SVM key metrics, depicted in Table 3, indicate that it makes better predictions for class 0 in terms of precision, but better for class 1 in terms of recall. The overall accuracy is moderate, that of 73%, and it suggests that the dataset might be a bit unbalanced.

Table 3. Classification report for SVM

Class	Precision	Recall	F1 score	Support
0	0.91	0.50	0.65	157
1	0.67	0.95	0.78	163

When we compare both results, then we see that XGBoost is better at avoiding false positives for Class 0 and SVM marginally better at avoiding false positives for Class 1. In both cases the model suggests that Class 0 or Gegë dialect is harder to predict, which might be also a case of noisy data.

Moreover, the classification report for RNN, presented in Table 4, suggests in general good performance, with no severe overfitting, and maybe suggests a minor imbalance in class 1.

Table 4. Classification report for RNN

Class	Precision	Recall	F1 score	Support
0	0.78	0.76	0.77	147
1	0.81	0.83	0.82	181

5. Conclusions

In conclusion, Recurrent Neural Networks (RNNs) again outperformed XGBoost and SVM in identifying dialectal variation by modelling the sequential dependencies of language more effectively.

A key limitation is our reliance on a small, manually annotated corpus (1,660 items) that may not capture slang, code-switching or spelling errors; forced Gegë⟶Toskë translations likely obscured natural dialect markers; heavy preprocessing (e.g., stripping punctuation) removed contextual cues; and subjective annotation—even with dictionary checks—could introduce bias, all of which constrain generalisability.

Future work should build a larger, more diverse social-media corpus; investigate transformer architectures (e.g., BERT and its multilingual variants); integrate richer linguistic features (morphology, syntax, phonetics); develop hybrid models combining sequential deep learning with traditional keyword-based methods; and validate their performance on real-world tasks such as sentiment analysis, topic modelling, and chatbots.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4o and Grammarly to improve the grammar and to spell check.

Acknowledgments

This research was supported by the Ministry of Education, Science, Technology and Innovation of Kosovo and HEI'25 project.

References

[1] J. Gjinari, Gj. Shkurtaj, Dialektologjia [Dialectology], ShBLU, Tiranë, 2000.

[2] L. Demiraj, The Origin of the Albanian: Language, History and Culture, 2008.

[3] F. Matzinger, The Albanian Language: Structure, History and Dialects, 2010.

[4] J. Gjinari, B. Bahri, Gj. Shkurtaj, Xh. Gosturani, Atlasi dialektologjik i gjuhës shqipe [Dialectological Atlas of Albanian Language], Vol. 1, Akademia e Shkencave e Shqipërisë, Instituti i Gjuhësisë dhe i Letërsisë; Università degli Studi di Napoli L'Orientale, Napoli, 2007.

[5] G. Meta, Challenges in Processing Albanian Dialects with NLP Techniques, Int. J. Comput. Linguist. 9(2) (2021) 56–72.

[6] B. Hoxha, Variability of Albanian Dialects in Online Communication, Albanian Linguistic Stud. J. 15(4) (2022) 123–140.

[7] M. Çeliku, The Toskë and Gegë Dialects: Linguistic Features and Historical Evolution, Academy of Albanian Studies, Tirana, 2005.

[8] S. Khaksar, A. Hassani, Ensemble Learning for Dialect Classification: A Case Study on Hawrami, Proc. of 2024 Int. Conf. Nat. Lang. Process., CEUR-WS.org/Vol-XXX/, 2024.

[9] N. Bariha, R. Sharma, J. Patel, Addressing Data Scarcity in Dialect NLP: The Binjhal Case Study, Lang. Resourc. Eval., 58(1) (2024) 85–102.

[10] O. Tibi, H. Messaoud, Arabic Dialect Identification Using Hamilton Neural Networks (HNN), J. Computat. Linguist., 42(3) (2025) 215–230.

[11] A. Shehu, A. Çomo, Using Twitter to Collect a Multi-Dialectal Corpus of Albanian, J. Lang Resourc., 15(2) (2023) 102–118.

[12] E. Trandafili, N. Kote, G. Plepi, Question Classification in Albanian Through Deep Learning Approaches, Int J Adv Comput Sci Appl., 14(3) (2023) 737–747.

[13] D. Parmar, H Bhavsar., Deep Ensemble Learning for Indo-Aryan Dialect Identification Procedia Computer Science225

[14] O. Zoubi, S.R. El-Beltagy, Arabic Dialect Identification on Social Media Using a Hybrid Transformer-Based Model, J. Nat. Lang. Eng. (2024).

[15] GegeToskeClassifiers Dataset and Algorithms, https://github.com/JetmirGjoni/GegeToskeClassifiers, 2025.

[16] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, Proc. of 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., CEUR-WS.org/Vol-XXX/, 2016.