ALBANIAN SENTIMENT ANALYSIS: EXPLORING MACHINE AND DEEP LEARNING TECHNIQUES

Alba HAVERIKU and **Elinda KAJO MEÇE**
Department of Computer Engineering, Faculty of Information Technology**,** Polytechnic University of Tirana, Albania
Corresponding author: alba.haveriku@fti.edu.al
AH, 0000-0003-2246-8408 **;** EKM, 0009-0002-7353-476

ABSTRACT

This study presents a comparative analysis of machine learning (ML) and deep learning (DL) algorithms for sentiment analysis in Albanian, a low-resource and morphologically rich language that poses distinctive challenges for natural language processing (NLP). This study explored the effectiveness of both traditional ML classifiers and neural network approaches, such as Long Short-Term Memory (LSTM) models. Experiments were conducted on two datasets: the AlbMoRe corpus and the Multilingual Twitter Sentiment Classification corpus. Given the complexity of Albanian, we examined how linguistic features influence algorithmic performance, noting that conventional ML models struggle to capture contextual sentiment effectively. We also performed a series of hyperparameter tuning experiments within the LSTM architecture, focusing on network depth, dropout rates, and embedding dimensions. These findings contribute to the advancement of NLP tools and resources for the Albanian language.

Keywords**:** sentiment analysis, machine learning, LSTM architecture, Albanian language.

1. INTRODUCTION

Albanian, a distinct branch of the Indo-European language family spoken by approximately 8 million people, is classified as a low-resource language in NLP (Hyllested and Joseph, 2022). Despite several isolated efforts to develop corpora and computational tools, most available resources remain limited in scope, proprietary, or lack comprehensive linguistic coverage. This scarcity hinders both research progress and practical NLP applications in this field.

Sentiment analysis, a prominent subfield of NLP, has gained considerable attention owing to its wide-ranging applications, including

marketing, social media monitoring, and customer experience analysis. While substantial research exists for high-resource languages such as English, French, and German, sentiment analysis for low-resource languages such as Albanian remains underexplored. The primary challenges are the limited annotated corpora, both in size and diversity, and the linguistic differences that distinguish Albanian from more widely studied languages.

This study evaluated the performance of various ML and DL algorithms for sentiment analysis tasks in Albanian. Two datasets were used in this study. First, the AlbMoRe corpus developed by Çano (2023) contains 800 manually annotated movie reviews labelled as positive or negative. Earlier experiments with this dataset (Çano 2023) applied traditional ML methods—Support Vector Machines (SVM), Logistic Regression, Decision Trees, and Random Forests—highlighting the need for further investigation using more advanced techniques.

The second dataset, compiled by Mozetič *et al*. (2016), includes over 1.6 million sentiment-annotated tweets in 13 languages, of which approximately 53,000 are Albanian. Using these two resources, we assessed the accuracy of both traditional and neural models. We also compared these results with performance on the English-language IMDb dataset (Maas *et al*. 2011), offering a cross-linguistic perspective.

The main contributions of this study are as follows:

1. A comparative evaluation of sentiment classification on both formal (movie reviews) and informal (tweets) Albanian corpora.
2. A cross-linguistic analysis comparing Albanian sentiment classification with English data, underscoring the challenges specific to low-resource languages.
3. A systematic comparison of traditional ML algorithms and DL models on low-resource data.

2. LITERATURE REVIEW: SENTIMENT ANALYSIS FOR ALBANIAN

Two of the most comprehensive corpora developed for the Albanian language are the Albanian National Corpus, created by researchers at Saint Petersburg State University (Arkhangelskiy, 2012), and the sqGlobe Corpus, developed at Beijing Foreign Studies University (Jing *et al*. 2012). The Albanian National Corpus contains approximately 16.6 million tokens, whereas sqGlobe comprises approximately 1 million words

annotated with part-of-speech tags and lemmatization. Despite their size and richness, both corpora are closed-source and thus unavailable for broader linguistic research purposes.

Within the Universal Dependencies (UD) initiative (Nivre *et al*. 2017), three treebanks currently exist for Albanian. The Treebank of Standard Albanian (TSA) consists of 60 sentences and 922 tokens annotated with part-of-speech and syntactic information (Toska *et al*. 2020). The Saarbrücken Treebank of Albanian Fiction (STAF) includes 202 sentences with 3,499 tokens (Talamo, 2025). The third, Gheg Pear Stories (GPS), was specifically built for the Gheg dialect and contains 966 sentences comprising 15,990 tokens (Ebert, 2022).

Kote *et al*. (2019) introduced a morphologically annotated corpus containing approximately 118,000 tokens. In a more recent contribution, Kote *et al*. (2024) presented the SALT (Standard Albanian Language Treebank), which consists of 1,400 sentences and 24,537 tokens. This treebank provides comprehensive annotations, including sentence and word segmentation, as well as Universal Part-of-Speech (UPOS) tags.

Having established a general overview of the major linguistic corpora available for Albanian, we now turn to the research on sentiment analysis and opinion mining. One of the earliest studies on sentiment analysis for Albanian was conducted by Biba and Mane (2014), who developed and evaluated a set of machine learning classifiers (SVM, Logistic Regression, Naïve Bayes) on a corpus of 400 political news articles. Their experiments reported accuracy rates between 86% and 92%, depending on the topics. Subsequently, Kote *et al*. (2018) presented an experimental setup comparing machine learning algorithms for opinion mining, achieving initial results ranging from 79% to 94% accuracy. In a later study, they trained and tested 50 different classification algorithms in Weka by combining data from five different topics.

Skënduli *et al.* (2018) implemented CNN-based models for emotion detection in Albanian microblogging texts, achieving accuracies above 90% on some subset corpora. Besjana *et al*. (2024) analyzed the sentiment of news articles and its role in predicting fake news. By including sentiment analysis as an attribute in their model, the authors improved the performance of the XGBoost model for fake news detection by 0.4%.

3. METHODOLOGY

This section outlines the methodological framework used to classify Albanian texts according to their expressed sentiments, categorizing each

instance as either positive or negative. The classification process comprised four main phases: i) data collection, ii) data preprocessing, iii) model building, and iv) model evaluation.

3.1. *Data collection & Pre-processing*

The first step in our pipeline involved collecting the data necessary for this study. We employed three corpora: two containing Albanian texts with sentiment annotations and one English-language corpus for comparative analysis. Table 1 presents the details of each dataset, encompassing both formal and informal language sources.

Table 1. Datasets used for model training and evaluation

Name of Corpus	Number of records	Language	Sentiment
AlbMoRe (Çano *et al.* (2023))	800 movie reviews	Albanian	positive, negative
Mozetic et al. (2016)	53,000 tweets	Albanian	positive, negative, neutral
IMDb (Maas *et al.* (2011))	25,000 movie reviews	English	positive, negative

The preprocessing phase is a critical step that ensures that raw text data are standardized and cleaned, making them suitable for feature extraction and model training. Our preprocessing pipeline involved several stages designed to remove noise and inconsistencies from the data set. Specifically, we applied the following procedures:

- **Stopword Removal:** Albanian stopwords were removed using a custom-built list to ensure that frequent but uninformative words did not dominate the feature space.
- **Noise Removal:** URLs and hashtags were removed from the datasets because they carried little semantic value relevant to sentiment.
- **Text Normalization:** All text was converted to lowercase to reduce dimensionality and improve consistency.
- **Label Filtering:** In the Mozetič *et al*. (2016) dataset, which originally contained three sentiment labels (positive, negative, and neutral), we excluded all records that were labeled as neutral. This allowed us to focus on a binary classification task that distinguished only between positive and negative sentiments.

- **Punctuation Removal:** All punctuation marks were removed from the datasets to simplify the feature space. Although punctuation can carry affective or stylistic cues relevant to sentiment, we opted for its removal in this study and suggested that future research explore punctuation as a potential feature for sentiment classification in Albanian.

This pipeline reflects a balance between linguistic sensitivity and computational efficiency, making it especially adaptable to low-resource languages, such as Albanian.

3.2. Model Building and Evaluation

In the model development phase, we evaluated both traditional machine learning algorithms and deep learning approaches using Long Short-Term Memory (LSTM) networks. For feature extraction, we applied the Term Frequency–Inverse Document Frequency (TF-IDF) technique, which quantifies the importance of words within a document relative to the corpus. TF-IDF was chosen for its computational efficiency, interpretability, and widespread use in sentiment analysis.

The processed datasets were split into training and testing subsets using an 80–20 ratio, with 80% of the records allocated for training and 20% for testing. We used the *random state* hyperparameter to control the randomness and ensure the reproducibility of the train–test splits across multiple executions.

Traditional machine learning models were implemented using the *Scikit-learn* library. For the LSTM architecture, we employed the *TensorFlow* and *Keras* frameworks. Initially, the LSTM model was tested using the default hyperparameters across all datasets. We then fine-tuned the LSTM model specifically for the AlbMoRe dataset, which yielded more consistent and accurate results. The tuned parameters included the: dropout rate in the LSTM layer, number of epochs, and batch size during model training. The impact of each hyperparameter adjustment on the model performance is presented in Table 1.

To assess the model performance, we used the *predict()* function from scikit-learn to generate predictions for the test data. The evaluation was based on two widely adopted metrics for classification tasks, accuracy and precision, which were selected to provide a balanced view of the models' performance. The results for each model are listed in Table 2.

4. RESULTS AND DISCUSSIONS

Table 2 presents the evaluation results for each algorithm across the different corpora. In this table, the terms “Corpus 1,” “Corpus 2,” and “Corpus 1+2” refer to the AlbMoRe, Mozetič *et al*. (2016) corpus, and their combined version, respectively.

Overall, the models performed best on the AlbMoRe corpus, achieving the highest accuracy and precision values among all the algorithms. In particular, the Support Vector Machine (SVM) algorithm obtained the highest accuracy of 94.4%. In contrast, models trained and tested on the Mozetič *et al*., (2016) corpus exhibited lower accuracy than those using the AlbMoRe and IMDb corpora. This reduction in performance may be attributed to less reliable sentiment annotations in the Albanian subset of the Mozetič corpus, which likely introduced noise and inconsistencies during training and evaluation. Additionally, the informal and short-text nature of social media content (e.g., Twitter) complicates accurate sentiment detection.

Notably, the performance of the models trained on Corpus 1+2 closely resembled that of Corpus 2 alone. This result can be explained by the large difference in size between the two corpora, whereby the substantially larger Mozetič corpus effectively dominates the combined dataset, thereby diminishing the influence of the smaller AlbMoRe corpus during training.

Table 2. Accuracy and precision metrics across algorithms and corpora

	SVM		Naïve Bayes		Logistic Regression		LSTM	
	Accuracy	Precision	Accuracy	Precision	Accuracy	Precision	Accuracy	Precision
Corpus 1	94.4%	0.94/0.95	93.1%	0.90/0.96	90.0%	0.89/0.91	91.3%	0.90/0.92
Corpus 2	79.6%	0.81/0.6	78.3%	0.78/0.73	78.5%	0.79/0.6	75.9%	0.84/0.48
Corpus 1+2	79.5%	0.81/0.66	77.9%	0.78/0.73	78.4%	0.79/0.73	76.0%	0.83/0.50
IMDb	89.4%	0.89/0.90	86.3%	0.87/0.85	89.1%	0.88/0.90	88.1%	0.87/0.89

Furthermore, we explored the effects of several hyperparameter settings on the LSTM model trained on the AlbMoRe corpus. The impact of each hyperparameter is reported in Tables 3-5, where only one parameter was varied at a time, while all others were held at their default values.

As shown in Table 3, increasing the batch size gradually improved the accuracy, with the highest accuracy (91.3%) achieved at a batch size of 128. Similarly, increasing the number of epochs (Table 4) produced consistent gains in performance, with the highest accuracy of 91.88% being reached after 10 epochs.

Table 5 shows the effect of dropout rate on LSTM accuracy. Dropout values of 0.1 and 0.4 yielded the highest observed accuracies (91.87% and 91.8%, respectively), whereas intermediate dropout values produced slightly lower performances. These results suggest that both smaller and larger dropout rates can help regularize the model without excessively diminishing its capacity to learn the patterns.

Overall, these findings demonstrate that both the algorithm choice and dataset quality exert a strong influence on sentiment classification performance. In particular, the AlbMoRe corpus supports more robust training outcomes, and careful hyperparameter tuning of the LSTM model can provide additional performance gains in deep learning–based sentiment classification tasks.

Table 3. Effect of Batch Size on LSTM Accuracy

Batch size	**16**	**32**	**64**	**128**
Accuracy (%)	**90.0**	**90.06**	**90.6**	**91.3**

Table 4. Effect of Number of Epochs on LSTM Accuracy

Number of Epochs	**5**	**6**	**8**	**10**
Accuracy (%)	**90.6**	**91.3**	**91.87**	**91.88**

Table 5. Effect of Dropout Rate on LSTM Accuracy

Dropout Rate	**0.1**	**0.2**	**0.3**	**0.4**
Accuracy (%)	**91.87**	**90.6**	**90.6**	**91.8**

5. CONCLUSIONS

This study presents a comparative analysis of several widely used machine learning algorithms, such as SVM, Naïve Bayes, and Logistic Regression, alongside an LSTM neural architecture for sentiment analysis in the Albanian language. Experiments were conducted across multiple corpora, including AlbMoRe, the Mozetič *et al*. (2016) dataset, and the English IMDb dataset, to evaluating model performance in terms of accuracy and precision.

The findings show that the AlbMoRe corpus consistently yielded the best performance, outperforming all other datasets. This result is largely attributed to the higher quality of annotation and well-structured sentences in AlbMoRe, as opposed to the less consistently labelled content found in the Mozetič dataset.

In addition to the baseline evaluations, hyperparameter tuning experiments for the LSTM model using the AlbMoRe dataset produced promising results. Optimized configurations of these parameters led to noticeable improvements in the classification performance. These findings highlight the potential of deep learning methods for Albanian sentiment analysis, emphasizing the importance of model optimization and dataset quality.

Future work may involve extensive hyperparameter searches and experimentation with more advanced deep learning architectures, such as Bidirectional LSTM, GRU, or transformer-based models such as BERT. Another important avenue is to increase the size and diversity of Albanian sentiment-related corpora by incorporating different dialects and domains (e.g., social media, literature, and news).

In conclusion, this study contributes to the growing body of research on Albanian as a low-resource language and provides evidence that, despite linguistic complexity and limited resources, acceptable performance can be achieved through careful corpus selection, algorithm testing and model fine-tuning.

DECLARATIONS

Declaration of AI use: We confirm that no AI-assisted technologies were used to create this article.

Originality and Authorship: This work is based on the original data and ideas developed by the authors.

Author contributions

Haveriku A: Ideation, conceptualization, methodology, data interpretation, resources, writing—original draft preparation, and corresponding author; **Kajo Meçe** E: Ideation, conceptualization, methodology, data interpretation, resources, review and editing, supervision.

Authors approval:

All authors have read and agreed to the published version of the manuscript.

Conflict of interest: The authors declare there are no conflicts of interest.

Funding: There is no related funding for this paper.

REFERENCES:

Arkhangelskiy T, Michael D, Yurievich RA, Alexandrovna MM. 2012. *Korpusi i gjuhës shqipe: Drejtimet kryesore të punës*. https://api.semanticscholar.org/CorpusID:185412758.

Biba M, Mane M. 2014. Sentiment analysis through machine learning: An experimental evaluation for Albanian. In: *Recent advances in intelligent informatics* (pp. 195–203). Springer.

Çano E. 2023. *AlbMoRe: Movie reviews in Albanian* [Corpus]. LINDAT/CLARIAH-CZ digital library, Charles University. http://hdl.handle.net/11234/1-5165.

Ebert C, Islamaj A, Kuqi A, Sonnenhauser B, Plamada M, Widmer P. 2022. UD Gheg Pear Stories [corpus].

Hyllested A, Joseph BD. 2022. Albanian. In J. Clackson (Ed.), The Cambridge handbook of Indo-European Languages (pp. 223–245). Cambridge University Press.

Jing K, Qiao J, Shihao Y, Tong H, Wang YF, Wang X, Hu Z, Lacmi E, Allmetaj E, Chen T, Zhang W, Zhang H, Lu Y, Ai W. 2012. The SQGlobe corpus: A balanced 1M-word contemporary written Albanian corpus, lemmatized and POS-tagged [corpus].

Kote N, Biba M, Kanerva J, Rönnqvist S, Ginter F. 2019. Morphological tagging and lemmatization of Albanian: A manually annotated corpus and neural models. arXiv preprint arXiv:1912.00991. http://arxiv.org/abs/1912.00991.

Skënduli MP, Biba M, Loglisci C, Ceci M, Malerba D. 2018. User-emotion detection through sentence-based classification using deep learning: A case-study with microblogs in Albanian. In: *International Symposium on Methodologies for Intelligent Systems* (pp. 258–267). Springer.

Kote N, Biba, M, Trandafili E. 2018. A thorough experimental evaluation of algorithms for opinion mining in Albanian. In International Conference on Emerging Internetworking, Data & Web Technologies (pp. 525–536). Springer.

Kote N, Rushiti R, Çepani A, Haveriku A, Trandafili E, Meçe Kajo E, Skënderi Rakipllari E, Xhanari L, Deda A. 2024. Universal Dependencies treebank for Standard Albanian: A new approach. In Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLiB 2024) (pp. 80–89). Institute for Bulgarian Language, Bulgarian Academy of Sciences. https://aclanthology.org/

Maas AL, Daly RE, Pham PT, Huang D, Ng AY, Potts C. 2011. Learning word vectors for sentiment analysis. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologi*es (pp. 142–150). Association for Computational Linguistics. https://aclanthology.org/P11-1015.

Mozetič I, Grčar M, Smailović J. 2016. *Twitter sentiment for 15 European languages* [corpus]. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1054.

Muraku B, Xiao L, Meçe EK. 2024. Toward detection of fake news using sentiment analysis for Albanian news articles. In L. Barolli (Ed.), Advances in Internet, Data & Web Technologies (EIDWT 2024), Lecture Notes on Data Engineering and Communications Technologies (Vol. 193, pp. 665–676). Springer. https://doi.org/10.1007/978-3-031-53555-0_55**Nivre J, Zeman D, Ginter F, Tyers F. 2017**. Universal Dependencies. In A. Klementiev & L. Specia (Eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*. Association for Computational Linguistics. https://aclanthology.org/E17-5001/.

Talamo L. 2025. Introducing STAF: The Saarbrücken Treebank of Albanian Fiction. Zenodo. https://doi.org/10.5281/zenodo.14552809.

Toska M, Nivre J, Zeman D. 2020. Universal dependencies for Albanian. In Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020). Association for Computational Linguistics.