Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 18 No. 24 (2024) | OPEN ACCESS

https://doi.org/10.3991/ijim.v18i24.50851

PAPER

AI-Based Hate Speech Detection in Albanian Social Media: New Dataset and Mobile Web Application Integration

Endrit Fetahi[1,2], Mentor Hamiti[1], Arsim Susuri[2](✉), Jaumin Ajdari[1], Xhemal Zenuni[1]

[1]South East European University, Tetovo, North Macedonia

[2]University of Prizren, Prizren, Kosovo

arsim.susuri@uni-prizren.com

ABSTRACT

This paper aims to advance AI-based hate speech (HS) detection in the Albanian language, which is resource-limited in natural language processing (NLP). Addressing the challenge of limited data, we developed a human-annotated dataset of over 11,000 comments, carefully curated from various Albanian social media platforms, containing a substantial number of HS instances. The dataset was annotated using a detailed two-layer taxonomy to capture the complex dimensions of HS. To ensure high-quality annotations, three expert annotators applied a majority voting system, achieving a substantial Fleiss's kappa coefficient of 0.62, underscoring the reliability and consistency of the annotations. We conducted a comparative analysis of several machine learning (ML) algorithms, including support vector machine (SVM), Naïve Bayes (NB), XGBoost, and random forest (RF), paired with various text vectorisation techniques and pre-processing methods. In binary classification, the NB model with term frequency-inverse document frequency (TF-IDF) vectorization achieved the highest performance, with an F1 score of 0.80. For multiclass classification, XGBoost outperformed other models, achieving an F1 score of 0.77. Interestingly, our experiments revealed that pre-processing steps generally reduced model performance, suggesting that raw text inputs work better for the Albanian language. Through error analysis using local interpretable model-agnostic explanations (LIME), we identified key challenges, such as polysemy and irony, which contributed to misclassifications. To demonstrate the practical applicability of our work, we developed a user-friendly mobile web application based on the best-performing model, providing real-time HS detection with the potential for integration into social media platforms.

KEYWORDS

hate speech (HS) detection, machine learning (ML), Albanian, social media networks, web platform

Fetahi, E., Hamiti, M., Susuri, A., Ajdari, J., Zenuni, X. (2024). AI-Based Hate Speech Detection in Albanian Social Media: New Dataset and Mobile Web Application Integration. *International Journal of Interactive Mobile Technologies (iJIM)*, 18(24), pp. 190–208. https://doi.org/10.3991/ijim.v18i24.50851

Article submitted 2024-07-01. Revision uploaded 2024-09-28. Final acceptance 2024-09-28.

1 INTRODUCTION

Over the last decade, the advancements in information technology have completely changed the way people interact. Online platforms that enable communication have largely replaced old-fashioned methods like in-person meetings and letters [1]–[3]. The rise of social media, messaging apps, and video conferencing capabilities has made it easier for people to collaborate remotely and share ideas [4], [5]. Social media, mostly accessed through smartphones, has shown to cause great addiction among youngsters and students [6]. Additionally, improvements in natural language processing (NLP) and machine learning (ML) have made it possible for people and machines to communicate more easily [7].

The technologically advanced era in which we currently live is not without its challenges. Unfortunately, the emergence of media platforms has made it easier for those with malicious motives to spread harming content, such as hate speech (HS) and disinformation, which is bad for both society and the individual [8]. Considering the amount of data generated on these platforms, most content shared there is neither reviewed nor controlled. As people increasingly communicate through these channels, phenomena like hate, offensive, and violent content are continuously on the rise [9]. This not only causes social discomfort but also erodes public trust in digital platforms, exacerbating the problem [10].

As a result of these problems, the need for automated systems that can identify hazardous content and stop its dissemination is increasing. NLP tools, enhanced by ML algorithms, have become essential to this effort. On the other hand, most of the techniques for detecting and reducing this kind of speech are only applicable to languages that have access to extensive hardware resources and massive annotated datasets [11]. Developing trustworthy solutions for low-resource languages is particularly difficult because of the obvious gap between high-resource languages and their technologically advanced counterparts [12].

This study aims to address these disparities by focusing on the identification of HS in a language with limited resources, specifically the Albanian language, with the goal of ensuring safety and inclusion on digital platforms. Closing this gap is crucial not only for equitable access to technology but also for promoting linguistic diversity. Research into techniques like transfer learning and multilingual training models may enhance the efficacy of NLP systems across various languages, leading to more widespread and inclusive digital engagement.

The main contributions of our study include:

- Developing a novel human-annotated HS dataset for the Albanian language through data collection and rigorous annotation processes.
- Implementing and evaluating various ML algorithms in different setups to detect and categorise hateful comments, considering the nuances of the language.
- Deploying the best-performing ML model as a web or standalone application, including a user-friendly mobile web interface and potential integration with social media platforms for real-time HS detection.

The research paper begins by analysing current developments in available datasets, both multilingual and within the Albanian context. The methodology of data scraping, and the annotation process are described in detail. ML algorithms are then implemented, and their results are discussed, including error analyses performed using explainable AI. Additionally, the development and deployment of a web application utilising the best-performing algorithm are highlighted, showcasing the practical application and scalability of the model. The outcomes of this study

open possibilities for further advancement in automated HS detection, specifically for the Albanian language. This could potentially lead to the development of more advanced and culturally sensitive technologies that better understand and mitigate harmful online behaviour.

2 RELATED WORK

Hate speech detection is a complex task that is heavily influenced by the quality of the dataset used. A high-quality dataset is crucial because it directly affects the performance and reliability of the detection models. Authors in the review [13] show that dataset quality, linguistic characteristics, and training properties of HS detection algorithms all affect the accuracy of such models. Therefore, datasets with balanced and representative samples enable models to learn effectively and generalise well to unseen data. There are many datasets for different languages, like English [14–16], German [17], Spanish [18], Italian [19], French [20], Dutch [21], Indonesian [22], Croatian [23], etc.

However, in the Albanian language, as shown in Table 1, existing datasets are limited in size and are unbalanced in terms of hate and non-hate instances. The Shaj Dataset [24] is one of the publicly available datasets in Albanian. This dataset provides unbalanced offensive instances, which can limit its effectiveness for training robust models. It adheres to the OffensEval [25] taxonomy for annotating comments. The authors of the dataset [26] discuss the initial efforts in automatic HS detection and classifiers for the Albanian language using binary classification based on the support vector machine (SVM) algorithm. Although the results appear promising, more extensive work is required to achieve comparative results, including expanding the HS corpus in size and further processing. They suggest that other ML models should be utilised and evaluated in future work. Another dataset for the Albanian language is mentioned in [27]. However, this dataset is based only on per-token (word) analysis and was aimed at developing a mobile system. The authors in [28] introduce a dataset for identifying abusive language, derived from scraped user comments on Kosovo social media channels on Facebook and YouTube. This dataset employs a three-layer OLID [25] annotation scheme, resulting in a total of 3,000 comments.

The human-annotated dataset we introduce aims to extend and enhance the quality of the final dataset, addressing the limitations of existing Albanian datasets. We achieve this by employing manual annotation conducted by multiple expert annotators to ensure high-quality and reliable labels. Additionally, we collect whole posts and focus on gathering as many hate comments as possible by selecting provocative posts that are likely to elicit HS. This approach not only increases the size of the dataset but also ensures a more balanced and representative sample of hate and non-hate instances, thereby improving the model's ability to detect HS effectively.

Table 1. Datasets available for the Albanian language

Source	Platform	Dataset	Origin	Language	# Size	# Classes
[24]	arXiv	Shaj	Instagram and Facebook	Albanian	10306 (normal) 1568 (offensive)	Hierarchy
[26]	GS	Zenuni et al.	Facebook	Albanian	4,886 total	Binary
[27]	IEEE	Raufi et al.	Social Media	Albanian	10,267 words	Binary
[28]	ACL	Ajvazi et al.	Facebook, YouTube, and Instagram	Albanian	2482 (normal) 518 (offensive)	Hierarchical

Machine learning methods are mostly employed for automatic HS detection. Authors in [29] developed a system using Naïve Bayes (NB) and SVM classifiers, programmed in Java and employing the Weka tools, with data from the UCI ML repository. When SVM classifiers were fine-tuned, they performed better than NB in identifying offensive content from non-offensive content. Meanwhile, authors in [30] show that random forest (RF) outperforms AdaBoost and neural network in the evaluation of three models on a HS detection dataset. In terms of predicted accuracy and dependability, RF outperformed both AdaBoost and neural network, achieving classification results with an F1 score of 71.3%. On the other hand, authors in [31] show that a model combining GloVe embeddings with a GRU network achieved the highest overall performance with a macro F1 score of 0.7680. In the multi-class classification problem, generally lower performance was observed across all models, with XGBoost leading at a macro F1 of 0.5871.

Deep learning approaches have been increasingly utilised in HS detection since they can automatically identify complicated patterns and representations from data. For example, the authors in [32] examined the use of deep neural network architectures and found that these methods performed significantly better than traditional techniques. The main advantages of deep learning approaches include their capacity to handle large datasets, capture intricate linguistic features, and improve the detection of context-dependent and subtle instances of HS, leading to more accurate and reliable models. In research [33], BERT in combination with CNN resulted in a high F1 score of 0.92, showing that these combinations are successful. In another study [34] for the Vietnamese language, Bi-LSTM compared to SVM, LR, and GRU showed the best performance with an F1 score of 71.43. The advantages of deep learning approaches include their ability to automatically learn complex patterns and representations from data, which can improve the detection of nuanced hate speech.

3 MATERIALS AND METHODS

3.1 Methodology

In this section we discuss the materials and procedures used throughout our research. Figure 1. shows the comprehensive workflow of our corpus design and experimental evaluation. While automatically scraping data from social media, we merge and clean the data. A two-layer system labels the data appropriately, and majority voting refines it, resulting in the final dataset. The final dataset has been curated and anonymised. We use many data preparation methods for analysis. Next, we structure the dataset for ML algorithms to categorise in different experimental setups. The final phase is a thorough model assessment to guarantee optimal performance.

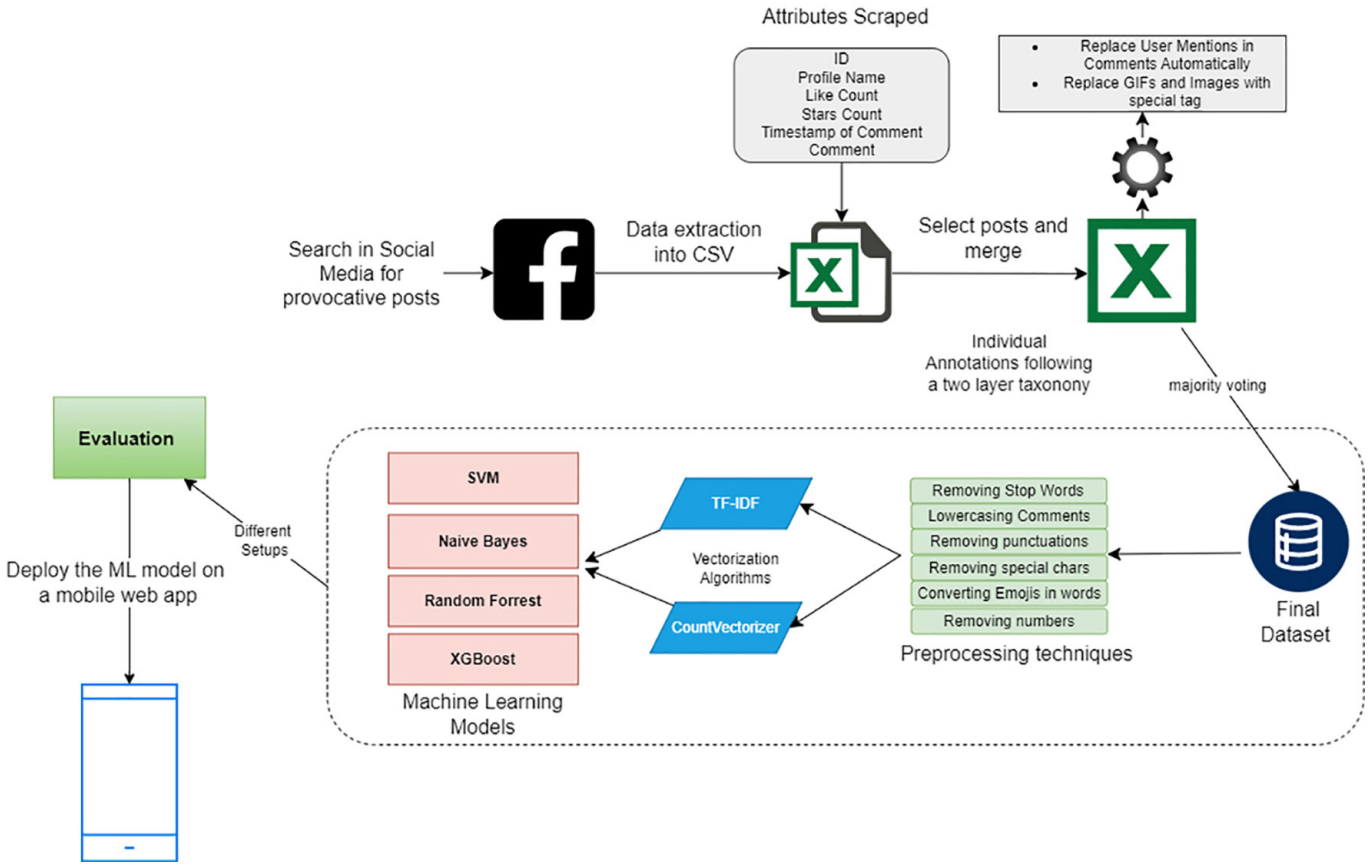

Fig. 1. Workflow of the data annotation and experiments

3.2 Data collection

The main source of the data gathered was the Facebook social network, as it is one of the most widely used social media sites in Albanian, with approximately 85% of the total population registered on it [35]. Also, it shows that in Kosovo, it is used almost by people of all ages [36].

The main tool to automatically scrape comments from Facebook was provided by the exportcomments.com tool. The procurement of data began with the identification of provocative and controversial posts in order to find as many HS comments as possible. During the crawling phase, in total, we identified and crawled around 84 posts. We chose the posts that had the highest number of comments. Currently, 11,343 comments were selected for annotation in order to align with the objective of the paper, which is to extract as many comments as possible from single posts. Due to the sensitive nature of the data and GDPR regulations, some personal information has been anonymized. The annotated dataset can be shared with researchers upon request, provided they follow data protection guidelines.

3.3 Annotation process

The quality and reliability of robust HS models are heavily dependent on the quality of the training data. For this reason, the annotation process of the proposed

dataset went through a rigorous continual process. Due to this, three highly qualified individuals were used for the annotation of the dataset in its labels. The three annotators are demographically from the region of Kosovo, fluent Albanian speakers, and highly educated with a background in computer science focused on natural language processing.

The categories that are used for labelling the comments follow a two-layer taxonomy. The foundation layer of our taxonomy presents a binary labelling process in which comments are initially passed to determine the presence or absence of HS elements. This binary classification is critical as it lays the groundwork for AI models in the future. Upon this process, nonetheless, it is then subjected to a more granular second-layer classicization. This layer elaborates on specific categories or nuances of HS in order for future work to extend the research.

The second layer consists of:

- General HS
- Religious HS
- Political HS
- Ethnicity HS

The classification has been meticulously developed to include all aspects of HS present in the cultural and social context. A sample of our dataset is shown in Figure 2.

Nonetheless, the annotators have been trained, and guidelines have been provided. Initially, the annotators have been guided to first annotate the first layer to determine if the comment contains HS or not, resulting in binary classification. After identifying the presence of HS, the annotators have been extending to the classes mentioned above.

In this study, we likewise use the term HS as an umbrella term for the different and numerous kinds of insulting user-created content, as mentioned by [37].

Nr	Post Nr	Reply Nr	Name	Profile ID	Date	Likes	Comment	HS Label	HS Category
1	1		[redacted]	[redacted]	24/10/22 [redacted]	0	Për hajr...	0	
2		1-1	[redacted]	[redacted]	24/10/22 [redacted]	0	Prap qasi jan mbrohen nepermjet Skudeve hu	1	4
3	2		[redacted]	[redacted]	24/10/22 [redacted]	2	ku e keni Kolonel ALBIN(veteran/dezerteri) kurto	1	3
4	3		[redacted]	[redacted]	24/10/22 [redacted]	8	Kujon veterant e rrejshum qi pomarin pare ntha	1	1
5	4		[redacted]	[redacted]	24/10/22 [redacted]	0	Veq ikan qit mej shetit	0	
6	5		[redacted]	[redacted]	25/10/22 [redacted]	0	shkojne siç vijne	0	

Fig. 2. Annotated dataset sample

3.4 Quality assurance

The quality and integrity of our dataset were maintained through a quality control process. The manual annotation process has been closely monitored, according to the given guidelines. For this purpose, quality control has been controlled by two aspects: the consensus of the labels achieved and the representation of the data.

Majority voting has been used to determine the final labels for the dataset, ensuring that the dataset is reliable. The Fleiss's Kappa coefficient is used to evaluate the reliability of the agreement among the annotators Fleiss's.

Table 2 displays the scoring convention utilized for the measurement of Fleiss' Kappa.

Table 2. Scoring convention of the Fleiss' Kappa results

< 0	Poor Agreement
0.01–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–1.00	Almost perfect agreement
< 0	Poor agreement

3.5 Experimental setup and evaluation

Machine learning algorithms. The algorithms that we will utilise includes SVM, NB, LR, and XGBoost. All experiments are conducted in a Python environment using well-known libraries such as sklearn [38], which include the above-mentioned algorithms and the text vectorizers.

- Support vector machine is a supervised ML technique used for classification and regression tasks. It works by identifying the hyperplane in the feature space that best divides the classes. The use of kernels in SVM allows the method to handle non-linear borders between classes, which makes it especially useful in high-dimensional environments.
- The NB classifier, on the other hand, is part of a family of probabilistic algorithms based on applying Bayes' theorem. Despite their simplicity, NB classifiers frequently exhibit good performance and are very scalable.
- LR is also a statistical method for predicting binary classes. The likelihood of a binary answer dependent on one or more predictor factors is modelled using an effective yet straightforward algorithm.
- XGBoost, an acronym for Extreme Gradient Boosting, is an efficient gradient-boosted decision tree implementation. Due to its success in several ML contests, this extremely complex algorithm has gained popularity. In terms of predictive power, XGBoost outperforms several other algorithms and offers a scalable and precise answer to a wide range of data science issues.

To ensure fairness and consistency, we used the default parameters for each classifier.

Pre-processing techniques. Before training the models, the dataset underwent a series of pre-processing techniques, which are essential for optimising the performance of the model. The pre-processing techniques used are:

- Removing stop words: A manually curated list of Albanian stop words is created by us and is removed in order to implement the experiments and enhance the overall accuracy of the models.
- Lowercasing: Converting all the text to lowercase
- Removing punctuation and special characters
- Removing numbers
- Converting emojis into standard words so that the model can understand the sentiment and nuances of the language

Text vectorisation. Text vectorisation algorithms are being implemented so text can be converted to numerical form in order for ML algorithms to understand it. The techniques we use are:

- Term frequency-inverse document frequency (TF-IDF) determines a word's importance within a document that is a part of a larger collection, or corpus. The more frequently a term appears in a given document, the more significant it is. This technique works well for separating common terms from those that are essential to understanding the context of texts. The computational formula of TF-IDF, in which "*t*" represents the terms, "*d*" refers to each individual document, and "*D*" stands for the entire collection of documents, is:

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D)$$

- Count vectorisation transforms text into a vector format, also known as a bag of words model. This paradigm ignores word order and grammar in favour of seeing a text as a collection of its constituent words. Text analysis is performed using ML models using the vectorised data that is produced. This method focuses just on the frequency with which a word occurs in the text.

Evaluation. To find the best-performing model, we review several key evaluation metrics that are crucial for assessing the performance of ML models. We use K-fold cross-validation to evaluate the classification models. It generates and tests numerous models on different dataset subsets, providing more accurate and less biased results in HS detection [31–33]. In our case of performance evaluation, we use 5-fold cross-validation. The various metrics we use to report in our experiments include accuracy, precision, recall, and the F1-score.

$$Accuracy = \frac{(Number\ of\ Correct\ Predictions)}{(Total\ Number\ of\ Predictions)}$$

$$Precision = \frac{(True\ Positives)}{(True\ Positives + False\ Positives)}$$

$$Recall = \frac{(True\ Positives)}{(True\ Positives + False\ Negatives)}$$

$$F1\ Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

Local interpretable model-agnostic explanations approach for error analysis. Machine learning models are particularly useful for automating HS detection content on social media. However, such models, when compiled, do not give much of an explanation, which can be a barrier for researchers and the experiments. In the research [39], it is shown that using explainable AI approaches such as LIME can be very useful, especially in HS detection.

Local interpretable model-agnostic explanations (LIME) is a technique used in explainable AI to show the predictions of ML models treating them as a black-box function [40]. It displays interpretable models to illustrate the process of making the predictions. Nevertheless, it is also model agnostic, showing that it can work with any model. It is being used in different AI systems to improve trust and show results that can be enhanced further.

Mobile web application. To facilitate the practical application of our model, we have also developed a simple mobile web application using the Python programming language as the backend, and HTML-CSS for the front end. We use the Flask web framework to handle the text input processing. For the prediction model, we use the best performing ML model based on our experiments. Using this setup, we effectively and efficiently do real-time text classification, integrating ML into a web application to aid users in detecting and understanding hate speech.

4 RESULTS

4.1 Dataset

This study introduces a novel dataset annotated for HS detection in the Albanian language, also known as a low-resourced language. The introduced dataset, to the best of our knowledge, is an Albanian dataset containing the highest number of hate comments compared to the existing datasets. The integrity and quality of the dataset were validated through a meticulous quality control process. We utilised Fleiss's Kappa to measure the level of agreement among the multiple expert annotators involved in manually labelling the dataset. Fleiss's Kappa measures the dependability of categorical rating agreement between a fixed number of ratters [41]. Unlike simple percentage agreement calculations, it accounts for the agreement occurring by chance, providing a more robust evaluation of inter-annotator reliability [42]. The consensus among annotators yielded a Fleiss's Kappa score of 0.62, indicating a substantial agreement level. This high level of agreement suggests that the labels are reliable, and that the dataset is of good quality. Ensuring such consistency in the annotation process is crucial in HS detection tasks because models heavily rely on accurately labelled data to learn effectively. By validating the dataset with Fleiss's Kappa, we enhance its credibility, which in turn positively affects the performance and generalisability of the detection models, improving the overall effectiveness of our HS detection system.

With the dataset's quality and reliability established, we show its composition to gain insights into the distribution of comments and HS categories. Based on the general statistics provided in Table 3, the annotated dataset shows a wide range of comments, most of which are neutral (54%) and fall into HS categories for the rest, demonstrating the variety of online communication. While the ratio for binary classification is relatively balanced, there are some differences in the distribution of HS classes. General HS is well represented, while religious, political, and ethnicity-based HS have fewer examples, with political HS being the most limited.

Table 3. Annotated dataset general statistics

	Neutral	General HS	Religious HS	Political HS	Ethnicity HS
# Of comments	6,131	3,901	598	148	565
Word count	86,710	41,852	8,500	2,163	9,080
Avg word per comment	14	10	14	14	16
Ratio (binary)	54.05%	45.94%			
Ratio (multiclass)	54.05%	34.39%	5.27%	1.30%	4.98%

4.2 Binary classification results

Several ML algorithms were compared, including SVM, NB, RF, and XGBoost. These were combined with two vectorisation techniques, TF-IDF and count vectorizer, with and without pre-processing, as shown in Table 4. The results indicate a consistent performance across the setups. NB in combination with TF-IDF showed the highest performance with an F1 score of 0.80. Comparable results were also shown by SVM with TF-IDF. However, RF and XGBoost showed slightly lower performance compared to the others.

Interestingly, by including pre-processing techniques, it demonstrated lower performance in all the models. This may imply that not processing the comments may have a positive impact, while also showing that the raw features captured by TF-IDF are already quite representative for the classification task for the Albanian language.

Table 4. Binary classification models

Classifier	Accuracy	Precision	Recall	F1
SVM + TF-IDF	0.79	0.79	0.79	0.79
SVM + TF-IDF + Pre-processing	0.78	0.78	0.78	0.78
SVM + CountVectorizer	0.76	0.76	0.76	0.76
SVM + CountVectorizer + Pre-processing	0.76	0.77	0.76	0.75
NB + TF-IDF	**0.80**	**0.80**	**0.80**	**0.80**
NB + TF-IDF + Pre-processing	0.79	0.79	0.79	0.79
NB + CountVectorizer	0.75	0.78	0.75	0.75
NB + CountVectorizer + Pre-processing	0.78	0.78	0.78	0.78
RF + TF-IDF	0.78	0.78	0.78	0.77
RF + TF-IDF + Pre-processing	0.77	0.78	0.77	0.76
RF + CountVectorizer	0.77	0.78	0.77	0.77
RF + CountVectorizer + Pre-processing	0.76	0.78	0.76	0.76
XGBoost + TF-IDF	0.77	0.77	0.77	0.76
XGBoost + TF-IDF + Pre-processing	0.75	0.77	0.75	0.74
XGBoost + CountVectorizer	0.77	0.78	0.77	0.77
XGBoost + CountVectorizer + Pre-processing	0.76	0.77	0.76	0.75

4.3 Multiclass classification results

For the multiclass classification, based on the results shown in Table 5, the performance metrics were generally lower compared to the binary classification results, which is expected due to the increased complexity of predicting HS among different classes. One of the primary reasons for the lower performance is the imbalance in the dataset across different labels. Some HS classes, like general HS, are more common, while others, such as political HS, are underrepresented. This imbalance makes it challenging for the models to learn and accurately predict the

less frequent classes, leading to a decrease in metrics like F1-score. Additionally, multiclass classification requires the model to distinguish between several categories rather than just two, increasing the difficulty of the task. The combination of label imbalance and the heightened complexity of differentiating between multiple classes contributes to the lower performance observed in the multiclass classification results.

Notably, XGBoost in all the setups demonstrated good results in the multiclass context, achieving an F1-score of 0.77 comparable to the binary classification setup. Nevertheless, in these experiments, including pre-processing techniques resulted in decreased performance for most of the algorithms. The other classification models performed poorly; while being the highest-performing model in the binary classification, the NB performed low in this case with a 0.73 F1 score in combination with CountVectorizer. However, we can note that the count vectorizer technique works better in the scenario of NB than with TF-IDF combinations. This indicates a potential trade-off between the simplicity and interpretability of models like NB in binary tasks and the complexity and nuanced understanding of models like XGBoost in multiclass tasks.

Table 5. Multiclass classification models

Classifier	Accuracy	Precision	Recall	F1
SVM + TF-IDF	0.77	0.80	0.77	0.69
SVM + TF-IDF + Pre-processing	0.77	0.76	0.77	0.68
SVM + CountVectorizer	0.76	0.66	0.76	0.68
SVM + CountVectorizer + Pre-processing	0.76	0.66	0.76	0.68
NB + TF-IDF	0.75	0.68	0.75	0.65
NB + TF-IDF + Pre-processing	0.75	0.68	0.75	0.65
NB + CountVectorizer	0.78	0.77	0.78	0.73
NB + CountVectorizer + Pre-processing	0.78	0.77	0.78	0.73
RF + TF-IDF	0.80	0.79	0.80	0.76
RF + TF-IDF + Pre-processing	0.79	0.78	0.79	0.75
RF + CountVectorizer	0.79	0.77	0.79	0.75
RF + CountVectorizer + Pre-processing	0.78	0.77	0.78	0.74
XGBoost + TF-IDF	**0.80**	**0.78**	**0.80**	**0.77**
XGBoost + TF-IDF + Pre-processing	**0.80**	**0.78**	**0.80**	**0.77**
XGBoost + CountVectorizer	**0.80**	**0.78**	**0.80**	**0.77**
XGBoost + CountVectorizer + Pre-processing	0.79	0.77	0.79	0.76

4.4 Error analysis

While having explainable AI approaches, we utilised LIME to critically analyse the performance of the best-performing model, in this case the NB with TF-IDF. In the following figures we show two examples where the model misclassified false positive and false negative and plotted the importance of the features.

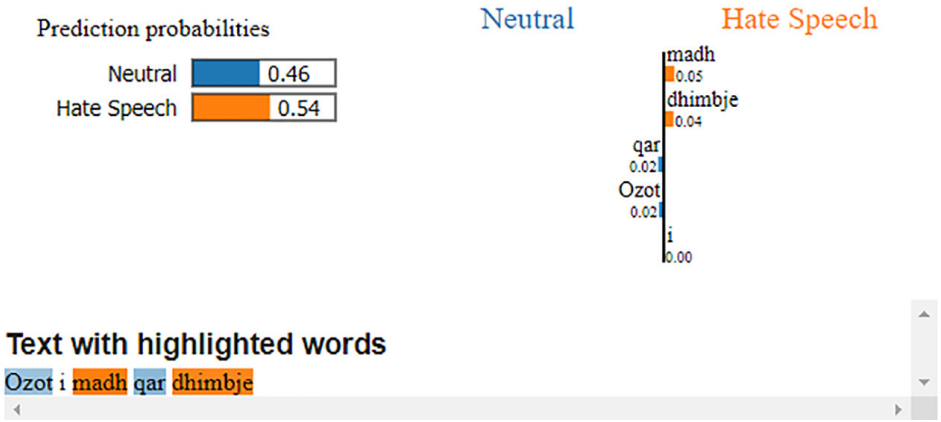

Fig. 3. LIME explanations for misclassified example 1

Example 1: False positive: In Figure 3, we present an instance where the model incorrectly labelled the sentence "Ozot I madh qar dhimbje" ("Oh Great God, what pain") as HS. Despite its benign intent, expressing sorrow or empathy, the model misclassified it as offensive. This misclassification highlights how polysemy—the presence of multiple meanings for a single word—can lead to errors. The word "dhimbje" ("pain") is used here to convey empathy but may be associated by the algorithm with contexts involving harm or offense, which are common in hate speech.

Implications: This example suggests that using contextualised language models or word embeddings—NLP techniques that take context into account, might improve the model. It also highlights how important it is to add a range of benign use examples of these terms to the training set to improve the models contextual discriminating skills.

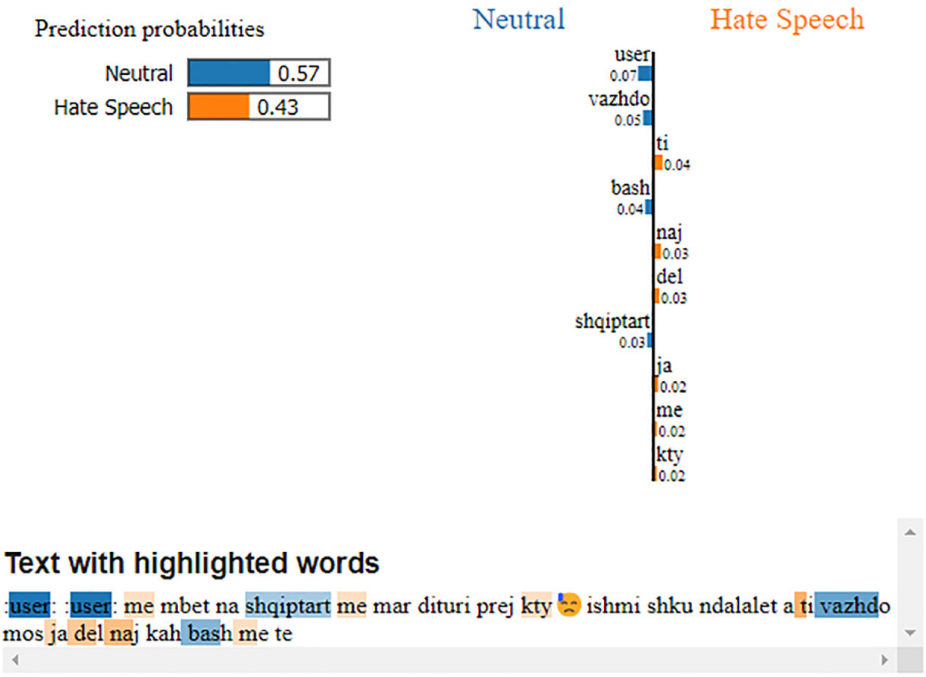

Fig. 4. LIME Explanations for misclassified example 2

Example 2: False negative: In Figure 4, we show a comment that the model failed to classify as HS, resulting in a false negative. The comment contains sarcastic and ironic language that conveys an unfriendly tone, but it is not overtly offensive in terms of explicit language. Because the algorithm relies heavily on surface-level indicators, such as particular offending phrases, it was unable to detect the underlying negative sentiment.

Implications: This incorrect classification highlights how the methodology is not able to identify implicit HS, sarcasm, or irony. To tackle this problem, sentiment analysis or sophisticated algorithms that can recognise complex negative phrases would need to be used.

4.5 Practical application

Web applications are software programs that run on a web server and are accessed by users through a web browser over a network like the Internet. They are designed to be used from any device with a web browser and an internet connection, eliminating the need for additional software installations. Their cross-platform compatibility, scalability, and ease of updating make them ideal for distributing ML models to a large audience.

For this study, we created a mobile web application with Python and HTML-CSS to demonstrate the real-world use of our HS detection model, which is presented in Figure 5. Frontend and backend technologies are integrated during the development of web applications to produce interactive platforms that are accessed via web browsers. In our instance, users interact with the software and receive feedback on whether the content contains HS through an intuitive interface and user-friendly service.

Our web application is written in HTML-CSS for the front end and Python for the back end, utilising the Flask framework. The NB Model with TF-IDF vectorisation, trained with scikit-learn, is loaded by the backend to handle the server-side logic and classify the input text. When a user enters text, the server processes the prediction result and sends it back to the frontend, where it is displayed to the user instantly. The user interface of the web application has been designed with an intuitive layout to facilitate user interaction.

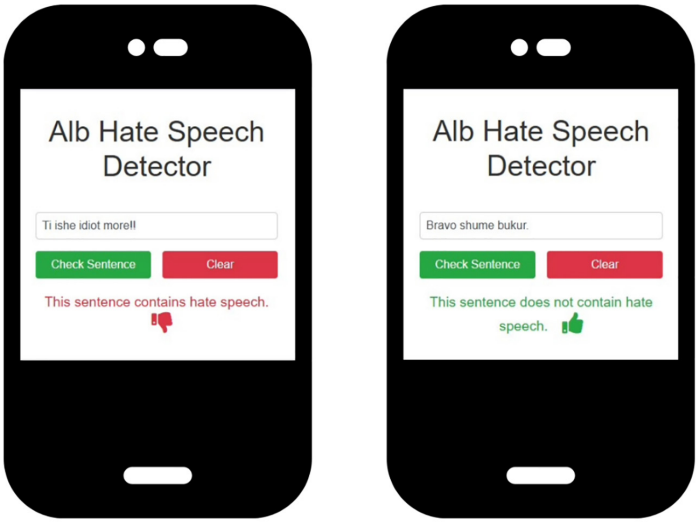

Fig. 5. Mobile web app using the deployed machine learning model

Despite our web application, the proposed ML model can be used as a stand-alone service for social media platform integration in addition to the web application. By facilitating the automatic, real-time detection and flagging of HS directly within the platforms, this enhances the model's scalability and usefulness. By immediately detecting and removing objectionable content, this integration improves the usefulness of the model and promotes safer online environments.

5 DISCUSSION

This study brings attention to the issue of automatically identifying HS on Albanian social media. We scraped social media data to create a comprehensive, human-annotated dataset with an appropriate balance of hate and non-hate instances. We further classified them into several HS nuances to provide additional value. The dataset was created under supervision and labelled by three expert annotators. The final labels were decided by majority vote. We used the Fleiss Kappa assessment to evaluate the dataset's quality and agreement; the result was a 0.62 coefficient, which suggests substantial agreement. In an effort to improve accuracy even further, we additionally annotated posts that we believed were provocative. Given that datasets have an important impact on models, this dataset acts as a foundation for further ML model development.

We used several types of ML models, including RF, XGBoost, NB, and SVM We employed both raw and pre-processed data to observe how well the models perform and to get ideas on what would work for the Albanian language, since other studies have utilized pre-processing. We employed the Count Vectorizer and TF-IDF techniques for feature extraction. With an F1 score of 0.80, the study's findings demonstrated that NB with TF-IDF produced the highest accuracy in binary classification. Though XGBoost outperformed the other algorithms in multiclass classification, its performance was still lower than that of binary classification. This is because models have difficulty distinguishing between different classes and we have imbalanced classes for the nuances of HS. It is likely that XGBoost performed better due to its tree-based structure, which allowed it to handle the features more effectively, while NB's simplicity made it more useful for binary classification. TF-IDF performed better than CountVectorizer because it could more successfully collect the most frequently occurring words.

Analysing the pre-processing techniques, we found that they didn't perform well for the Albanian language. One factor is the lack of good NLP tools like PoS tagging, stemming, and lemmatisation to proceed with pre-processing. Additionally, the language used in social media comments is highly deviated, resulting in non-standard language, which poses a problem for pre-processing. We conducted an error analysis using LIME and observed that the misclassified comments were often due to polysemy and irony. The model needs more exposure to such comments in the training set.

5.1 Strengths and weaknesses of the approach

Among the study's strengths is the introduction of a carefully developed dataset for the detection of HS in Albanian, which was thoroughly evaluated by several annotators to increase its reliability. It is also based on a real-world case scenario. We demonstrated the usefulness of this data set by using it to further develop an

AI-based system. We found the best strategies for Albanian HS identification by experimenting with different algorithms and feature extraction techniques. Yet there are also specific weaknesses in the approach. In multiclass classification in particular, the models have trouble properly distinguishing the nuances of HS. This restriction comes in part from the unequal handling of nuances in HS and the deviated language used on social media, which frequently consists of multiple dialects and lacks grammatical structure. Albanian's lack of advanced NLP tools limited the effectiveness of pre-processing methods. Furthermore, both annotation and model training were limited by the difficult nature of the dataset development, where many comments were unreadable and language-deviated.

5.2 Comparing with existing research

In Table 6, we show a comparison of various research efforts in HS detection in Albanian, focusing on dataset size, best performing models, accuracy, and relevant characteristics. The table highlights important differences in dataset size and balance, both of which seem to play a significant role in model performance. For instance, our approach, using a larger and nearly balanced dataset, achieves a strong F1 score of 0.80. This contrasts with other studies, such as one using SVM, which reported a lower F1 score of 0.58, possibly due to the smaller dataset size. While a BERT model reached a higher F1 score of 0.86, it relied on a small and imbalanced dataset, which may limit its general applicability.

Table 6. Comparison of the existing literature

Source	Language	Dataset	# Size	Best Performing Model	Best Accuracy	Characteristics
[24]	Albanian	Shaj	10306 (normal) 1568 (offensive)	BERT	0.77 ACC	– Highly imbalanced – Paper not peer-reviewed
[26]	Albanian	Zenuni et al.	4,886 total	SVM	0.58 F1	– No clear guidelines on the annotation process – Smaller dataset
[28]	Albanian	Ajvazi et al.	2482 (normal) 518 (offensive)	BERT	0.86 F1	– Smaller dataset – Highly Imbalanced
Our approach			6131 (normal) 5212 (hate)	NB + TF-IDF	0.80 F1	– Carefully annotated – Real-World Utility – Multiple expert annotators – Nearly Balanced – Preprocessing impact analysing

5.3 Practical implications of the findings

We have carefully developed HS detection models using AI techniques to aid the community in advancing this field. The results of our study can serve as a baseline or be extended in future work. By showcasing the practical use of such models through a web application, we contribute to creating safer online environments. The developed web app showcases the suitability of our models and offers a showcase for incorporating HS identification into platforms, improving efforts towards moderation in Albanian.

5.4 Future research directions

For future work, we suggest

- Exploring transfer learning models, such as Transformer architectures, which could enhance the HS detection process
- Developing suitable NLP tools for more advanced pre-processing of the Albanian texts
- Developing domain-specific word embeddings to better capture the linguistic nuances of the Albanian language
- Expanding the dataset through crowdsourcing with a larger and more diverse population could improve its representativeness and quality, since perceptions of offensive content can vary greatly among individuals

6 CONCLUSION

This study implements different ML algorithms to detect HS on Albanian social media. These models were trained on a novel HS corpus which was carefully crafted and annotated. Based on the extensive number of experiments conducted, we conclude that detecting HS for the Albanian language, despite its limited NLP resources, is possible with good accuracy using the proposed methods. It was the TF-IDF features that aided the ML models to perform better on raw comments than pre-processing the text at all. This was due to the lack of advanced NLP tools for pre-processing available for the Albanian language. The experiments proved that the simplicity of NB performed better for detecting HS in the binary setup, while in the multiclass scenario, tree-based methods like XGBoost could determine the labels better. Through error analysis, it was revealed that in Albanian social media, people tend to offend using irony and sarcasm, which even for the ML models was hard to recognize. Polysemy also showed to be challenging for detection of HS. Additionally, the study findings emphasise the need for the development of further NLP tools for the Albanian language. Extending the data even further to a large-scale scenario, with plentiful data in the training set, would probably enhance the detection system. The significance of this study advances the field in terms of scientific value by introducing a new dataset and offering practical solutions and further insights into HS detection for the Albanian language.

7 REFERENCES

[1] T. Tülübaş, T. Karakose, and S. Papadakis, "A holistic investigation of the relationship between digital addiction and academic achievement among students," *Eur. J. Investig. Heal. Psychol. Educ.*, vol. 13, no. 10, pp. 2006–2034, 2023. https://doi.org/10.3390/ejihpe13100143

[2] Z. Van Veldhoven and J. Vanthienen, "Digital transformation as an interaction-driven perspective between business, society, and technology," *Electron. Mark.*, vol. 32, pp. 629–644, 2022. https://doi.org/10.1007/s12525-021-00464-5

[3] M. S. Albulayhi and S. El Khediri, "A comprehensive study on privacy and security on social media," *Int. J. Interact. Mob. Technol. (iJIM)*, vol. 16, no. 1, pp. 4–21, 2022. https://doi.org/10.3991/ijim.v16i01.27761

[4] J. A. N. Ansari and N. A. Khan, "Exploring the role of social media in collaborative learning the new domain of learning," *Smart Learn. Environ.*, vol. 7, 2020. https://doi.org/10.1186/s40561-020-00118-7

[5] W. Wagino, H. Maksum, W. Purwanto, W. Simatupang, R. Lapisa, and E. Indrawan, "Enhancing learning outcomes and student engagement: Integrating e-learning innovations into problem-based higher education," *Int. J. Interact. Mob. Technol. (iJIM)*, vol. 18, no. 10, pp. 106–124, 2024. https://doi.org/10.3991/ijim.v18i10.47649

[6] T. Karakose, T. Tülübaş, and S. Papadakis, "Revealing the intellectual structure and evolution of digital addiction research: An integrated bibliometric and science mapping approach," *Int. J. Environ. Res. Public Health*, vol. 19, no. 22, p. 14883, 2022. https://doi.org/10.3390/ijerph192214883

[7] V.-D. Păvăloaia and S.-C. Necula, "Artificial intelligence as a disruptive technology—A systematic literature review," *Electronics*, vol. 12, no. 5, p. 1102, 2023. https://doi.org/10.3390/electronics12051102

[8] A. Dreißigacker, P. Müller, A. Isenhardt, and J. Schemmel, "Online hate speech victimization: Consequences for victims' feelings of insecurity," *Crime Sci.*, vol. 13, 2024. https://doi.org/10.1186/s40163-024-00204-y

[9] M. Khalafat, J. S. Alqatawna, R. Al-Sayyed, M. Eshtay, and T. Kobbaey, "Violence detection over online social networks: An arabic sentiment analysis approach," *Int. J. Interact. Mob. Technol. (iJIM)*, vol. 15, no. 14, pp. 90–110, 2021. https://doi.org/10.3991/ijim.v15i14.23029

[10] X. Zhou *et al.*, "Hate speech detection based on sentiment knowledge sharing," in *ACL-IJCNLP 2021 – 59th Annu. Meet. Assoc. Comput. Linguist. 11th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, 2021, pp. 7158–7166. https://doi.org/10.18653/v1/2021.acl-long.556

[11] S. S. Aluru, B. Mathew, P. Saha, and A. Mukherjee, "Deep learning models for multilingual hate speech detection," *arXiv preprint arXiv:2004.06465*, pp. 1–16, 2020. https://doi.org/10.48550/arXiv.2004.06465

[12] T. Ranasinghe and M. Zampieri, "Multilingual offensive language identification for low-resource languages," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 21, no. 1, pp. 1–13, 2022. https://doi.org/10.1145/3457610

[13] E. Fetahi, M. Hamiti, A. Susuri, V. Shehu, and A. Besimi, "Automatic hate speech detection using natural language processing: A state-of-the-art literature review," in *2023 12th Mediterranean Conference on Embedded Computing (MECO)*, 2023, pp. 1–6. https://doi.org/10.1109/MECO58584.2023.10155070

[14] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the type and target of offensive posts in social media," *NAACL HLT 2019*, vol. 1, pp. 1415–1420, 2019. https://doi.org/10.18653/v1/N19-1144

[15] J. Pavlopoulos, J. Sorensen, L. Laugier, and I. Androutsopoulos, "SemEval-2021 Task 5: Toxic Spans Detection," in *SemEval 2021 – 15th Int. Work. Semant. Eval. Proc. Work.*, 2021, pp. 59–69. https://doi.org/10.18653/v1/2021.semeval-1.6

[16] T. Caselli, V. Basile, J. Mitrovic, I. Kartoziya, and M. Granitzer, "I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language," *Lr. 2020 – 12th Int. Conf. Lang. Resour. Eval. Conf. Proc.*, 2020, pp. 6193–6202.

[17] C. Demus, J. Pitz, M. Schütz, N. Probol, M. Siegel, and D. Labudde, "DeTox: A comprehensive dataset for German offensive language and conversation analysis," in *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, 2022, pp. 143–153. https://doi.org/10.18653/v1/2022.woah-1.14

[18] M. Álvarez-Carmona *et al.*, "Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets," in *CEUR Workshop Proc.*, 2018, vol. 2150, pp. 74–96.

[19] C. Bosco, M. Sanguinetti, F. Dell'Orletta, F. Poletto, and M. Tesconi, "Overview of the EVALITA 2018 hate speech detection task," in *CEUR Workshop Proc.*, 2018, vol. 2263. https://doi.org/10.4000/books.aaccademia.4503

[20] A. Ollagnier, E. Cabrio, S. Villata, and C. Blaya, "CyberAgressionAdo-v1: A dataset of annotated online aggressions in French collected through a role-playing game," in *2022 Lang. Resour. Eval. Conf. Lr.*, 2022, pp. 867–875.

[21] T. Caselli *et al.*, "DALC: The Dutch Abusive Language Corpus," in *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, 2021, pp. 54–66, 2021. https://doi.org/10.18653/v1/2021.woah-1.6

[22] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, "Hate speech detection in the Indonesian language: A dataset and preliminary study," in *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2017, pp. 233–238. https://doi.org/10.1109/ICACSIS.2017.8355039

[23] R. Shekhar, M. Karan, and M. Purver, "CoRAL: A context-aware croatian abusive language dataset," in *2nd Conf. Asia-Pacific Chapter Assoc. Comput. Linguist. 12th Int. Jt. Conf. Nat. Lang. Process. – Find. Assoc. Comput. Linguist. AACL-IJCNLP 2022*, 2022, no. 2021, pp. 217–225.

[24] E. Nurce, J. Keci, and L. Derczynski, "Detecting Abusive Albanian," *arXiv preprint arXiv:2107.13592*, 2021.

[25] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "SemEval-2019 Task 6: Identifying and categorizing offensive language in social media (OffensEval)," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019, pp. 75–86. https://doi.org/10.18653/v1/S19-2010

[26] J. Ajdari, F. Ismaili, B. Raufi, and X. Zenuni, "Automatic hate speech detection in online contents using latent semantic analysis," *Pressacademia Procedia*, vol. 5, no. 1, pp. 368–371, 2017. https://doi.org/10.17261/Pressacademia.2017.612

[27] B. Raufi and I. Xhaferri, "Application of machine learning techniques for hate speech detection in mobile applications," in *2018 Int. Conf. Inf. Technol. InfoTech 2018 – Proc.*, 2018, pp. 1–4. https://doi.org/10.1109/InfoTech.2018.8510738

[28] A. Ajvazi and C. Hardmeier, "A dataset of offensive language in Kosovo social media," in *2022 Lang. Resour. Eval. Conf. Lr. 2022*, 2022, pp. 1860–1869.

[29] D. C. Asogwa, C. I. Chukwuneke, C. C. Ngene, and G. N. Anigbogu, "Hate speech classification using SVM and Naive BAYES," *IOSR Journal of Mobile Computing & Application (IOSR-JMCA)*, vol. 9, no. 1, pp. 27–34, 2022.

[30] K. Nugroho *et al.*, "Improving random forest method to detect hatespeech and offensive word," in *2019 Int. Conf. Inf. Commun. Technol. (ICOIACT 2019)*, 2019, pp. 514–518. https://doi.org/10.1109/ICOIACT46704.2019.8938451

[31] R. Kumar, V. Gupta, and R. Pamula, "Hate speech and offensive content identification in English tweets," in *CEUR Workshop Proc.*, 2021, vol. 3159, pp. 104–109.

[32] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *26th Int. World Wide Web Conf. 2017 (WWW 2017 Companion)*, 2017, no. 2, pp. 759–760. https://doi.org/10.1145/3041021.3054223

[33] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A BERT-based transfer learning approach for hate speech detection in online social media," in *Complex Networks and Their Applications VIII, COMPLEX NETWORKS 2019, Studies in Computational Intelligence*, H. Cherifi, S. Gaito, J. Mendes, E. Moro, and L. Rocha, Eds., vol. 881, pp. 928–940, 2020. https://doi.org/10.1007/978-3-030-36687-2_77

[34] H. T.-T. Do, H. D. Huynh, K. Van Nguyen, N. L.-T. Nguyen, and A. G.-T. Nguyen, "Hate speech detection on Vietnamese social media text using the bidirectional-LSTM model," *arXiv preprint arXiv:1911.03648*, pp. 4–7, 2019. https://doi.org/10.48550/arXiv.1911.03648

[35] S. Kemp, "DIGITAL 2021: KOSOVO," Datareportal, 2021. [Online]. Available: https://datareportal.com/reports/digital-2021-kosovo

[36] Hallakate, "Online users in Kosovo by age," 2022. [Online]. Available: https://hallakate.com/en/online-users-in-kosovo-by-age/

[37] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 2017, pp. 1–10. https://doi.org/10.18653/v1/W17-1101

[38] G. Hackeling, *Mastering Machine Learning with scikit-learn*. 2014. [Online]. Available: http://books.google.com/books?id=fZQeBQAAQBAJ&pgis=1

[39] H. Mehta and K. Passi, "Social media hate speech detection using explainable artificial intelligence (XAI)," *Algorithms*, vol. 15, no. 8, p. 291, 2022. https://doi.org/10.3390/a15080291

[40] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, and W. Samek, "Explainable AI Methods – A Brief Overview," in *xxAI – Beyond Explainable AI, xxAI 2020, Lecture Notes in Computer Science,* A. Holzinger, R. Goebel, R. Fong, T. Moon, K. R. Müller, and W. Samek, Eds., vol. 13200, 2022, pp. 13–38. https://doi.org/10.1007/978-3-031-04083-2_2

[41] K. L. Gwet, "Large-sample variance of Fleiss generalized Kappa," *Educ. Psychol. Meas.*, vol. 81, no. 4, pp. 781–790, 2021. https://doi.org/10.1177/0013164420973080

[42] F. Moons and E. Vandervieren, "Measuring agreement among several raters classifying subjects into one-or-more (hierarchical) nominal categories. A generalisation of Fleiss' kappa," *arXiv preprint arXiv:2303.12502*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.12502

8 AUTHORS

Endrit Fetahi is a full-time teaching assistant at the University of Prizren. He holds a bachelor's and master's degree in computer science and is currently a PhD candidate at the South East European University (E-mail: ef30456@seeu.edu.mk).

Mentor Hamiti is a full-time professor at the Faculty of Contemporary Sciences and Technologies, South East European University in Tetovo, Macedonia.

Arsim Susuri is an Associate Professor at the University of Prizren 'Ukshin Hoti.' He holds a Ph.D. in Computer Science (E-mail: arsim.susuri@uni-prizren.com).

Jaumin Ajdari is a full-time professor at the Faculty of Contemporary Sciences and Technologies at South East European University in Tetovo, Macedonia.

Xhemal Zenuni is a full-time professor at the Faculty of Contemporary Sciences and Technologies at South East European University in Tetovo, Macedonia.