DOI: 10.2478/seeur-2024-0025

INTEGRATING HANDCRAFTED FEATURES WITH MACHINE LEARNING FOR HATE SPEECH DETECTION IN ALBANIAN SOCIAL MEDIA

Endrit Fetahi, Mentor Hamiti, Arsim Susuri, Xhemal Zenuni, Jaumin Ajdari

Faculty of Contemporary Sciences and Technologies
South East European University,
Tetovo, North Macedonia
ef30456@seeu.edu.mk, m.hamiti@seeu.edu.mk, xh.zenuni@seeu.edu.mk,
j.ajdari@seeu.edu.mk

Faculty of Computer Sciences
Uninversity of Prizren Ukshin Hoti, Prizren, Kosovo
arsim.susuri@uni-prizren.com

ABSTRACT

Online social media has seen a significant increase in usage over the last decade, enabling people to communicate more easily. The vast amount of data generated by these platforms is mostly uncontrolled and unmanageable. This has also provided opportunities for individuals to engage in hate speech and offensive language on these platforms. To address this issue, this research aims to conduct extensive experiments using machine learning models and handcrafted feature extraction in the low-resource language Albanian. We utilized several machine-learning algorithms, including Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), and Logistic Regression (LR), and extracted a considerable number of handcrafted features. To improve accuracy, we carefully performed feature selection to identify the most relevant features for detecting hate speech in the Albanian language. The results show that LR performed best in terms of accuracy, with an F1 score of 76.77. Using Random Forest feature ranking and SHAP analysis revealed that many comments on Albanian social media exhibit unique characteristics, resulting in a large feature set. This suggests that there is no clear pattern for the machine learning models to accurately flag the comments, indicating that Albanian is linguistically challenging to analyze.

Key words: Hate speech detection, Machine learning, handcrafted features, Albanian, social media.

INTRODUCTION

In recent years, the significant development of information technologies has altered the way people communicate. In particular, the rise of online social media has exponentially increased the number of users who use these portals daily to express their opinions and engage in debates. While this growth appears positive, it has led to a surge of unregulated data, often lacking proper tools for management. This has allowed threat actors to engage in harmful behavior on these platforms (Del Vigna et al., 2017). Hate speech and offensive language have been increasing in many countries, and this rise has been associated with various related crimes.

The rise in uncontrolled content has prompted researchers to develop automated systems for detecting harmful speech (Turki & Roy, 2022). Recent progress in natural language processing (NLP) and machine learning has demonstrated potential in overcoming these obstacles (Ayo et al., 2020; Beyhan et al., 2022; Khairy et al., 2021). The wide variety of terms used in hate speech, from slang and emojis to linguistic deviations that frequently contradict basic grammar rules, makes hate speech detection a challenging task. This has shown especially to be a challenge in languages which are limited in NLP tools. Even though there is limited work done on developing tools and extracting handcrafted features for Albanian (Canhasi et al., 2022; Fetahi et al., 2024; Misini et al., 2024), such features can enhance machine learning performance by capturing contextual details specific to Albanian language.

Several studies have demonstrated promising, highly accurate hate speech detection systems in high resourced languages like English (Alharthi et al., 2023; Mozafari et al., 2020), or Spanish (Álvarez-Carmona et al., 2018).

However, limited research has been conducted on hate speech detection in the Albanian language. (Ajdari et al., 2017) took initial steps in developing an automated hate speech classifier for Albanian texts using the SVM algorithm. Initial results are promising, with an F1 score of 0.58, but they highlight the need for a richer hate speech corpus and more advanced language processing features for Albanian to improve performance. In another study, (Nurce et al., 2021) investigate both classical machine learning and deep learning models using their established Albanian dataset for offensive speech detection. They provide a public dataset for this purpose, and their experiments with the BERT transformer model achieved an accuracy of 0.77.

Machine learning and deep learning models have demonstrated remarkable efficacy in hate speech detection, particularly through their ability to learn complex patterns from large datasets (Fortuna & Nunes, 2019). These models performance can be further improved by incorporating handcrafted features (Reddy, 2024).

The research describes how combining handcrafted features with machine learning models can improve the accuracy of hate speech classification, especially in complicated social media settings where language usage is extremely context-dependent and varied (Nascimento et al., 2023).

The main objective of this research is to investigate how hate speech detection systems can be strengthened by combining machine learning methods with a comprehensive set of handcrafted features, including diverse linguistic attributes and contextual information

specifically tailored to data scraped from real-world platforms. The work presents an effective solution for identifying and mitigating hate speech on Albanian online platforms, offering results as a baseline for future NLP improvements.

This research follows a logical and experimental flow, beginning with a thorough data collection and annotation process. Expert annotators labeled each comment, and majority voting was used to determine the final labels. Furthermore, a considerable number of handcrafted features were manually extracted and analyzed for the feature selection process. Finally, extensive machine learning model training was conducted using the selected feature subsets, and the results are thoroughly analyzed and discussed.

DATA COLLECTION

The dataset and its quality are fundamental for machine learning models to function properly (Fetahi et al., 2023). A high-quality dataset will enhance the model in automating the process in a proper manner. In this section, we describe the details of our dataset, which is currently under development. The dataset we use in this research paper and for its experiments is human-annotated and represents a real-world utilization case. We have automatically scraped Facebook comments from Albanian social media. The identification of the posts was manually done through different keywords and pages that have many subscribers and where people tend to comment frequently. In order to obtain as many hate comments as possible, we were influenced to select provocative posts, which would trigger the audience to comment. This was done because most public datasets, even in high-resource languages, lack a balance of hate speech and non-hate speech documents. The dataset underwent several phases:

- Raw comments were scraped from different public Facebook pages to obtain a considerable number of comments.
- The comments were further cleaned to remove names, which were replaced with the token user, and URLs were converted into a single token.
- Each comment was annotated by three expert annotators to ensure a high-quality dataset.
- The annotations and their labels were further validated by the inter-annotator Fleiss Kappa agreement, resulting in a 0.63 value, indicating substantial agreement and good annotation.
- We used majority voting to determine the final label.
- Upon completion of the dataset, we proceeded with coding the methods for automatically extracting features for each comment, resulting in a matrix.
- We experimented with feature selection methods to interpret and find the best-performing model under these circumstances.
- Each model was further trained with the features we extracted and were evaluated in terms of accuracy.
- The results were interpreted and discussed.

The dataset consists of 15,947 comments in total, and its distribution is shown in Table 1. Figure 1 illustrates that most comments are short in length, with the majority containing up to 40 words per comment. Additionally, the most frequent words are displayed in Figure 2.

Table 1. Annotated Dataset size

	Non-Hate speech	Hate speech
Number of comments	9,000	6,947

Number of words	110,627	124,477

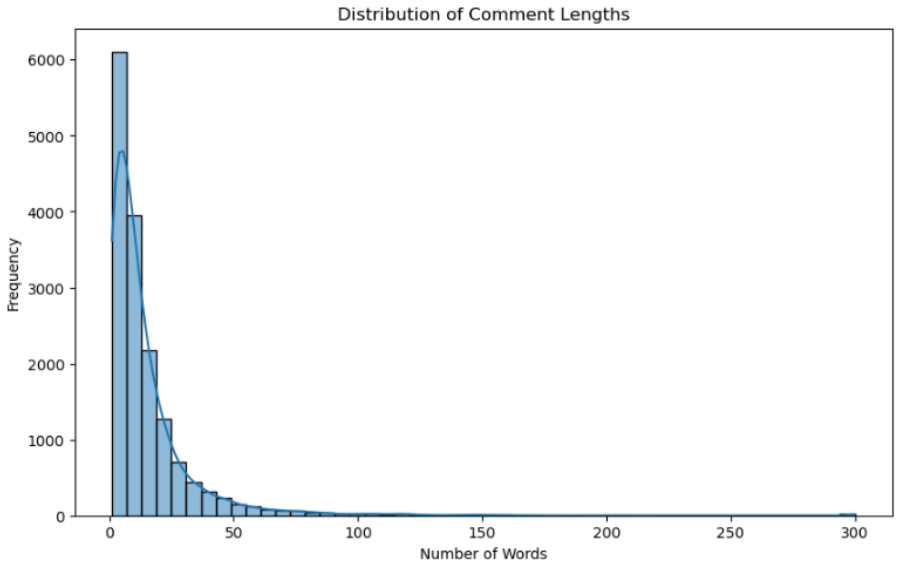

Figure 1. Distribution of comments lengths based on number of words

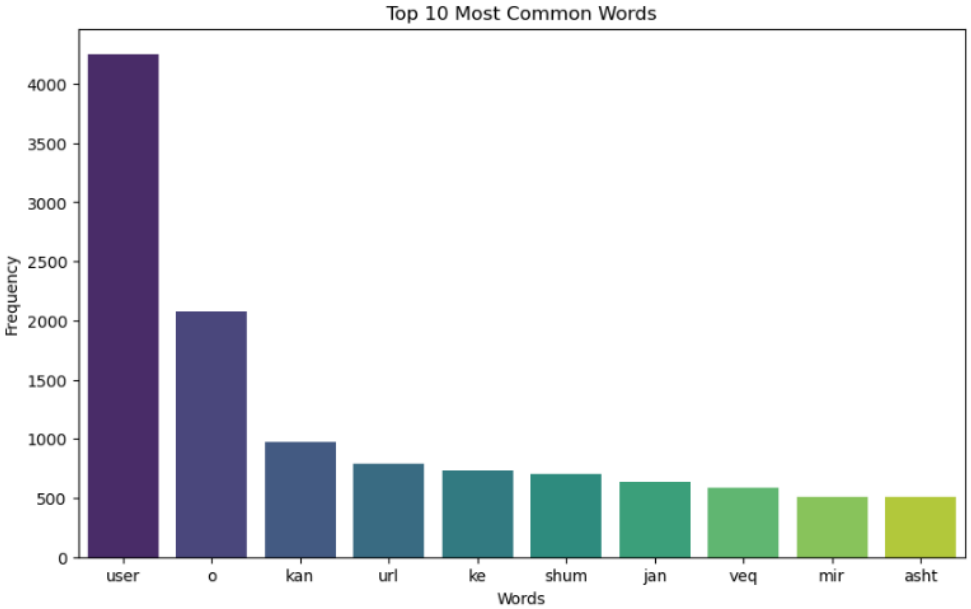

Figure 2. Most frequent words on the dataset

MACHINE LEARNING MODELS

To obtain high accuracy and interpretability in hate speech detection, we intend to apply handcrafted features in combination with different ML models. Each of the models was executed on a Python environment using scikit library (Hackeling, 2014). We experimented on four distinct ML algorithms:

- SVM (Support Vector Machine)
- Random Forest (RF)
- Naïve Bayes (NB)

- Logistic Regression (LR)

To ensure a reasonable and fair comparison of their performance on unseen data, all of these models were trained and tested using an identical train/test split. As is common, we used a 75% training and 25% testing split to ensure that we have enough data for both training and testing, which allows us to do thorough evaluations. Using the same random seed for data splitting allows us to keep things consistent and repeatable across studies. We did employ the default hyper parameters for the models.

HANDCRAFTED FEATURE EXTRACTION

In order to achieve good experimental results, we did extract a considerable number of features for each comment. We did manually code each of the functions using a Python environment. The final number of features extracted for each comment is 834 features. However, this has not come without its challenges. Since this research is focused on the Albanian language, it is highly lacking in NLP (Natural Language Processing) tools. Even so, for some of the methods, we manually curated lists where we thought it was possible, such as a stop word list, positive and negative word lists, conjunctions, prepositions lists, etc. We acknowledge that these are not official and can be further extended in the future by professional linguists. A summary of the features we have extracted is shown in Table 2.

Table 2. Handcrafted feature list

Feature Category	Features Descriptions
Textual Metrics	Word and character counts, sentence structure, readability, and lexical diversity
Punctuation and Symbols	Frequency and types of punctuation marks, special characters, and their usage patterns
Capitalization Patterns	Distribution of uppercase, lowercase, and capitalized words
Vocabulary Diversity	Variety of unique words, lexical richness, and word complexity
Repetition Metrics	Instances of repeated characters, words, and punctuation
N-gram Analysis	Frequency and average lengths of unigrams, bigrams, trigrams, and character-based n-grams
Phonetic Characteristics	Syllable counts, consonant and vowel ratios
Structural Layout	Document structure including lines, paragraphs, and formatting elements
Sentence Structure	Types and lengths of sentences, including declarative, interrogative, and exclamatory
Sentiment Indicators	Counts of positive and negative terms
User Engagement Metrics	Interaction statistics such as replies, likes, and user activity patterns
Emoji and Hashtag Usage	Presence and diversity of emojis and hashtags
URL Presence	Inclusion and frequency of URLs within the text
Controversial Content Indicators	Placement and frequency of potentially sensitive or controversial terms
Language Dynamics	Detection of language switching or code-mixing within the text
Semantic Similarity Measures	Similarity scores comparing text to predefined sets of offensive and positive content
Total Features extracted: 834	

FEATURE SELECTION AND METHODS

When it comes to how well machine learning models perform, not all extracted features have equal significance. We must separate and retain the most significant features if our hate speech detection algorithm is to work as efficiently as possible.

We opted to use the RF model as our feature selection strategy because of its tree-based architecture, which makes it ideal for ranking the relevance of features. The feature selection process involves the use of three approaches:

1. **Whole Feature Set** – In this approach we evaluate the machine learning models using every feature we have extracted. It is a solid starting point, giving us a baseline to compare against other feature selection techniques. This shows the impact of throwing every feature in the training process, even them that may be irrelevant or insignificant. Its value lies in revealing how the model performs when all possible features are considered, helping us understand the influence of adding even the less important ones.
2. **Feature Selection Based on Feature Importance with Ranking Threshold -** Random Forest's natural structure in which it is developed, makes prioritize features based on their relevance. At each split, the Random Forest's decision trees prioritize features according to how effectively they decrease impurity. We can reduce the size of the feature set by using a threshold, in which in our case, we used a 0.001 threshold, which could lead to better model performance and easier interpretability, also removing the features with no impact. This method is popular since it eliminates extra information or features that don't provide any value.
3. **Recursive Feature Elimination (RFE) –** A widely recognized method that works by iteratively removing the least important features, aiming to identify the subset of features that contribute most to the model's performance (Ramezan, 2022). Using this approach, not only can we increase accuracy, but it also helps reduce overfitting and remove redundant features, which can enhance the model's ability to detect hate speech. In our experiments, due to hardware constraints, we used a step size of 50 for selecting features incrementally. As a result, we show the number of features and their importance using the above methods.

The RF model was chosen for feature analysis due to its ability to effectively rank and interpret feature importance, particularly by prioritizing features with higher interpretability, even though some other algorithms may perform better in terms of accuracy (Bénard et al., 2022; Orlenko & Moore, 2021).

EVALUATION METRICS AND INTERPRETABILITY

In order to thoroughly evaluate our models, we use measures like Accuracy, F1 score, Precision, and Recall. With these measures, we can see how well we are doing at detecting hate speech while also keeping the number of false positives and negatives to a minimum.

We also use SHAP (SHapley Additive exPlanations) for visualizations and prioritize the features to make them easier to understand. SHAP is a method to explain the results produced by ML models (Chen et al., 2022). According to Shapley values, which are derived from game theory, each feature or feature value should be given credit for the model's prediction (Lundberg & Lee, 2017). With this method, the decision-making process may be better understood by seeing how each feature affects the model's predictions.

EXPERIMENTAL RESULTS AND DISCUSSION

In this chapter, we present the results of the hate speech detection experiments conducted using the methods described in the previous chapters. The findings are organized with the performance metrics and comparative model evaluations. The results are displayed in Table 3.

The experimental results show that when utilizing the full set of features, the Random Forest model performs the best, achieving an F1 score of 76.51. Additionally, its precision and recall are well balanced. It is further followed by LR with an F1 score of 70.54, however, RF still outperforms LR. In contrast, NB and SVM exhibit lower performance.

When applying the feature ranking with thresholding technique, there is a considerable improvement in the case of Naive Bayes, with a 5% increase in F1 score. However, RF and LR experience minor drop in performance, suggesting that these models are better suited to handling a larger feature set. Feature reduction through ranking appears to have a slightly negative impact on their effectiveness, while SVM performance remains nearly unchanged.

The results for Recursive Feature Selection reveal that LR benefited the most from this method, resulting as the best-performing model across all experiments, with an F1 score of 76.77. This suggests that LR succeeds on a more concise and relevant subset of features. RF also showed a slight improvement, while NB experienced a drop in performance, likely due to its sensitivity to feature independence assumptions. SVM in this subset of features also continued to perform poorly, demonstrating no improvement across any of the feature selection methods.

Overall, RF and LR resulted as the most promising models for hate speech detection in this task, particularly when paired with the appropriate feature selection techniques.

Table 3. Experimental results for the ML models on different feature sets

Feature Selection Method	Model	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
Whole set of Features	**Naive Bayes**	74.30	66.11	64.97	66.11
	SVM	63.06	63.56	62.90	63.56
	Random Forest	76.49	76.52	**76.51**	76.52
	Logistic Regression	71.25	70.40	70.54	70.40
Ranking Feature with Thresholding Technique	**Naive Bayes**	72.40	71.66	70.46	71.66
	SVM	63.08	63.58	62.92	63.58
	Random Forest	76.31	76.40	**76.32**	76.40
	Logistic Regression	71.18	70.60	70.73	70.60
Recursive Feature Selection	**Naive Bayes**	69.28	67.62	67.74	67.62
	SVM	63.08	63.58	62.92	63.58
	Random Forest	76.53	76.57	76.54	76.57
	Logistic Regression	76.91	76.98	**76.77**	76.98

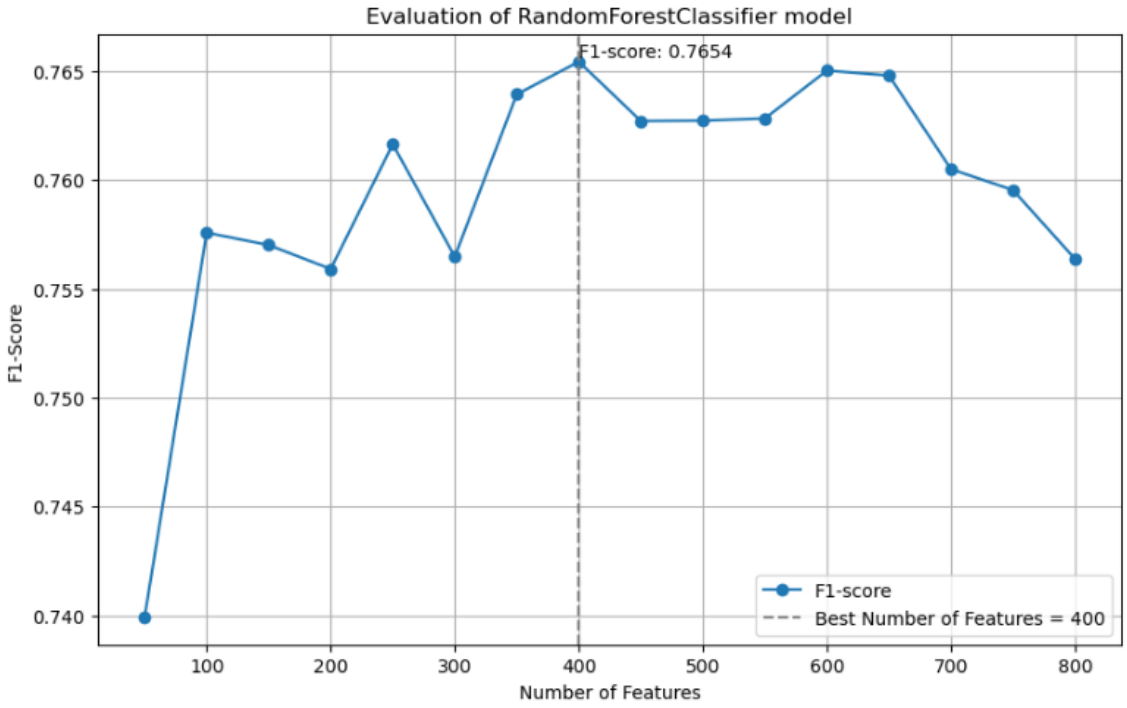

Figure 3. RFE using Random Forest and step size = 50

Figure 3 shows the F1-score evaluation of the RF classifier model as a function of the number of features used. With an increase of features, the model's performance increases gradually and reaches its highest at 400 features, with an F1-score of 0.7654. This suggests that to fully capture the complexity found in social media comments written in Albanian, more features may be necessary.

It also shows us that individuals often post comments in a variety of ways, using complex word and expression patterns that make it challenging for the model to detect hate speech based on a smaller feature set or identifying a clear pattern for the detection. It appears that adding more features does not always make the model any better at identifying hate speech, since the model's performance reaches an upper limit and then starts to reduce after 400 features.

The difficulties in processing comments in Albanian are highlighted by this finding, as there are few distinct linguistic patterns, and the model needs a lot of features to handle the linguistic nuances.

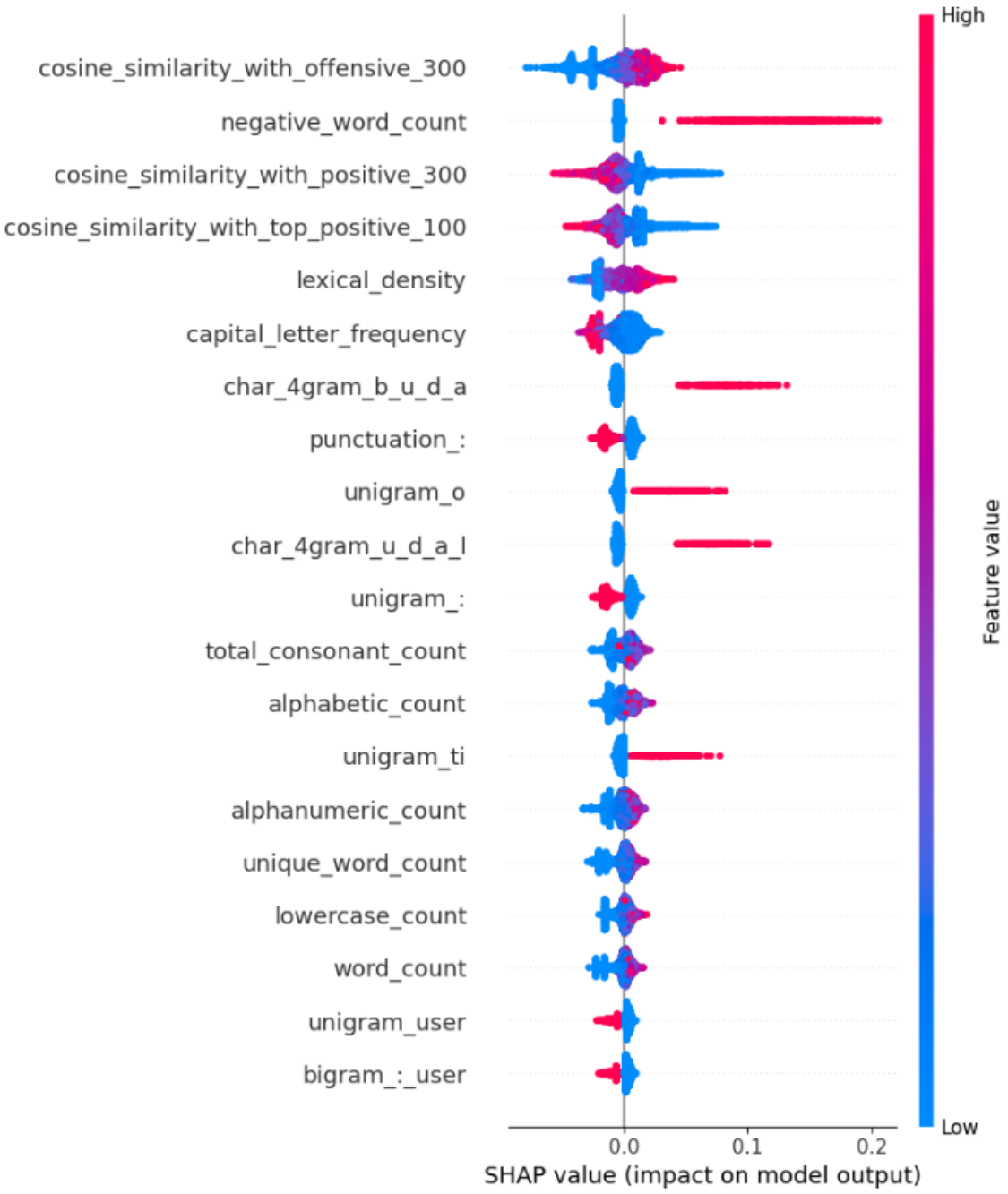

Figure 4. SHAP summary plot using Random Forest Classifier and RFE feature set

The SHAP plot in Figure 4. shows the most important features that affect the RF model results in detecting hate speech on Albanian social media. The most significant feature is cosine similarity with a document containing 300 offensive comments. A higher similarity level makes it more likely that hate speech will be found, which shows that the cosine similarity with that document affects the ability to detect HS. Low similarity with documents with positive comments is another way to use its effect to detect hate speech. On the other hand, a high negative word count is strongly linked to harmful comments. The number of words used, and the amount of capital letters aid the predictions furthermore. This shows that short, content-packed language and heavy capitalization often indicate hostile speech. The model is affected by character-level n-grams, punctuation patterns, and linguistic factors such as the number of

consonants. This shows how complicated Albanian language patterns are. The fact that user-specific terms are present also indicates that the way individuals address other people by tagging them in comments plays a role in detecting hate and offensive language.

The feature ranking generated by RF shown in Table 4., demonstrates the difficulty of identifying hate speech as no single element stands out as being the most important, having equal importance spread across features. Important factors like negative word count, lexical density, n-gram length, capital letter frequency, and cosine similarity with offensive and positive terms score highly. This shows that semantic content and stylistic elements are significant. Since no single pattern can adequately capture the wide variety of hate speech expressions, especially in the Albanian language, hate speech detection requires a broad and nuanced approach. This is evidenced by the diversity of features that have identical feature values across all features.

Table 4. Feature ranking and value generated by RF

Rank	Feature name	Feature Value	Rank	Feature name	Feature Value
1	negative_word_count	0.0274	11	alphanumeric_count	0.0125
2	cosine_similarity_with_offensive_300	0.0233	12	total_consonant_count	0.0121
3	cosine_similarity_with_positive_300	0.0195	13	avg_unigram_length	0.0119
4	lexical_density	0.0190	14	text_length_bytes	0.0118
5	cosine_similarity_with_top_positive_100	0.0179	15	consonant_to_vowel_ratio	0.0115
6	capital_letter_frequency	0.0169	16	var_word_length	0.0112
7	unique_word_count	0.0143	17	std_dev_word_length	0.0112
8	avg_word_length_variability	0.0132	18	cosine_similarity_with_top_offensive_100	0.0111
9	avg_bigram_length	0.0126	19	vowel_ratio	0.0109
10	avg_trigram_length	0.0125	20	avg_word_length	0.0106

CONCLUSION

In this study, we developed and evaluated different machine learning models in combination with handcrafted features to provide insights into automatic hate speech detection which happens in Albanian social media portals. Our study demonstrates that detecting hate speech in Albanian is a complex task, requiring a significant number of features for accurate detection. We employed a diverse set of 834 handcrafted features across different machine learning models such as Random Forest, Logistic Regression, Support Vector Machine, and Naïve Bayes. While we used the entire feature set, we also carefully performed feature selection, creating two subsets, one by using Recursive Feature Elimination and the other by Feature Selection Based on Feature Importance with Ranking Threshold. With the full feature set, the best-performing model was RF; however, when using RFE, the subset of features worked better in LR, achieving the highest F1 score of 76.77%. Feature ranking of RF and

SHAP analysis revealed that no single feature dominates, showing that each comment in Albanian social media has its own unique characteristics and user behavior towards hate speech, which requires a large number of features for achieving the best F1 score. This conclusion is further strengthened by the feature ranking of RF, which shows that each feature has a nearly equal effect on the model's performance Moreover, a combination of content, form, and context contributes to accurate classification. For future work, we propose exploring the use of transfer learning models, such as transformer architectures, and further developing advanced NLP tools, as Albanian lacks resources, to extract richer features.

REFERENCES

1. Ajdari, J., Ismaili, F., Raufi, B., & Zenuni, X. (2017). Automatic hate speech detection in online contents using latent semantic analysis. *Pressacademia*, *5*(1), 368–371. https://doi.org/10.17261/pressacademia.2017.612
2. Alharthi, R., Alharthi, R., Shekhar, R., & Zubiaga, A. (2023). Target-Oriented Investigation of Online Abusive Attacks: A Dataset and Analysis. *IEEE Access*, *11*, 64114–64127. https://doi.org/10.1109/ACCESS.2023.3289148
3. Álvarez-Carmona, M., Guzmán-Falcón, E., Montes-y-Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., Reyes-Meza, V., & Rico-Sulayes, A. (2018). Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets. *CEUR Workshop Proceedings*, *2150*, 74–96.
4. Ayo, F. E., Folorunso, O., Ibharalu, F. T., & Osinuga, I. A. (2020). Machine learning techniques for hate speech classification of twitter data: State-of-The-Art, future challenges and research directions. *Computer Science Review*, *38*, 100311. https://doi.org/10.1016/j.cosrev.2020.100311
5. Bénard, C., Veiga, S. Da, & Scornet, E. (2022). Interpretability via Random Forests. In *Interpretability for Industry 4.0 : Statistical and Machine Learning Approaches* (pp. 37–84). Springer International Publishing. https://doi.org/10.1007/978-3-031-12402-0_3
6. Beyhan, F., Çarık, B., Arın, İ., Terzioğlu, A., Yanikoglu, B., & Yeniterzi, R. (2022). A Turkish Hate Speech Dataset and Detection System. *Proceedings of the Language Resources and Evaluation Conference*, *June*, 4177–4185. https://aclanthology.org/2022.lrec-1.443
7. Canhasi, E., Shijaku, R., & Berisha, E. (2022). Albanian Fake News Detection. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *21*(5), 1–24. https://doi.org/10.1145/3487288
8. Chen, H., Lundberg, S. M., & Lee, S.-I. (2022). Explaining a series of models by propagating Shapley values. *Nature Communications*, *13*(1), 4512. https://doi.org/10.1038/s41467-022-31384-3
9. Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on Facebook. *CEUR Workshop Proceedings*, *1816*(January), 86–95.
10. Fetahi, E., Hamiti, M., Susuri, A., Selimi, B., & Saiti, D. I. (2024). Neural Network and Transformer-Based PoS Tagger for Low Resource Languages. *2024 International Conference on Information Technologies (InfoTech)*. https://doi.org/10.1109/InfoTech63258.2024.10701401
11. Fetahi, E., Hamiti, M., Susuri, A., Shehu, V., & Besimi, A. (2023). Automatic Hate Speech Detection using Natural Language Processing: A state-of-the-art literature review. *2023 12th Mediterranean Conference on Embedded Computing (MECO)*, 1–6. https://doi.org/10.1109/MECO58584.2023.10155070
12. Fortuna, P., & Nunes, S. (2019). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, *51*(4), 1–30. https://doi.org/10.1145/3232676
13. Hackeling, G. (2014). Mastering Machine Learning with scikit-learn. In *Book*. http://books.google.com/books?id=fZQeBQAAQBAJ&pgis=1
14. Khairy, M., Mahmoud, T. M., & Abd-El-Hafeez, T. (2021). Automatic Detection of Cyberbullying and Abusive Language in Arabic Content on Social Networks: A Survey. *Procedia CIRP*, *189*, 156–166. https://doi.org/10.1016/j.procs.2021.05.080
15. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model

predictions. *Advances in Neural Information Processing Systems*, *2017-Decem*(Section 2), 4766–4775.
16. Misini, A., Canhasi, E., Kadriu, A., & Fetahi, E. (2024). Automatic authorship attribution in Albanian texts. *PLOS ONE*, *19*(10), e0310057. https://doi.org/10.1371/journal.pone.0310057
17. Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). Hate speech detection and racial bias mitigation in social media based on BERT model. *PLoS ONE*, *15*(8 August), 1–26. https://doi.org/10.1371/journal.pone.0237861
18. Nascimento, F. R. S., Cavalcanti, G. D. C., & Da Costa-Abreu, M. (2023). Exploring Automatic Hate Speech Detection on Social Media: A Focus on Content-Based Analysis. *SAGE Open*, *13*(2). https://doi.org/10.1177/21582440231181311
19. Nurce, E., Keci, J., & Derczynski, L. (2021). Detecting Abusive Albanian. *ArXiv Preprint ArXiv:2107.13592*.
20. Orlenko, A., & Moore, J. H. (2021). A comparison of methods for interpreting random forest models of genetic association in the presence of non-additive interactions. *BioData Mining*, *14*(1), 9. https://doi.org/10.1186/s13040-021-00243-0
21. Ramezan, C. A. (2022). Transferability of Recursive Feature Elimination (RFE)-Derived Feature Sets for Support Vector Machine Land Cover Classification. *Remote Sensing*, *14*(24), 6218. https://doi.org/10.3390/rs14246218
22. Reddy, A. N. (2024). Enhancing Hate Speech Detection with Integrated Content-Based and Stylistic Features. *J.ElectricalSystems*, 3660–3666.
23. Turki, T., & Roy, S. S. (2022). Novel Hate Speech Detection Using Word Cloud Visualization and Ensemble Learning Coupled with Count Vectorizer. *Applied Sciences (Switzerland)*, *12*(13). https://doi.org/10.3390/app12136611