Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 19 No. 20 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijim.v19i20.56307

PAPER

Challenges and Solutions in Clustering Low-Resource Language Social Media Text: An Evaluation Using Unsupervised Algorithms

Mërgim H. Hoti, Avni Rexhepi (✉), Arbër H. Hoti, Blerim Rexha

University of Prishtina, Prishtinë, Republic of Kosova

avni.rexhepi@uni-pr.edu

ABSTRACT

Low-resource languages present unique challenges for natural language processing (NLP) due to limited annotated corpora, linguistic resources, and pre-trained models. This paper addresses the gap in clustering methodologies for such languages by evaluating the performance of three unsupervised algorithms—K-Means, DBSCAN, and HDBSCAN—on social media text data. Unlike prior studies focusing on high-resource languages, this study explores challenges in preprocessing, tokenization, and vectorization specific to low-resource settings. The results highlight the sensitivity of clustering performance to linguistic nuances and preprocessing approaches, with DBSCAN and HDBSCAN excelling in handling noisy and unstructured data. The findings provide actionable insights into algorithm selection and preprocessing strategies, showcasing the potential and limitations of traditional clustering methods in low-resource NLP. By shedding light on these challenges, this study paper contributes to the development of inclusive approaches for text analysis across underrepresented languages, advancing NLP applications globally.

KEYWORDS

unsupervised algorithms, K-Means, DBSCAN, HDBSCAN, and low-resource language

1 INTRODUCTION

Low-resource languages such as Albanian present unique challenges for natural language processing (NLP) due to the scarcity of annotated corpora, linguistic resources, and pre-trained models. While substantial progress has been made in NLP for high-resource languages such as English, these advancements do not always translate to low-resource contexts.

Albanian presents several linguistic characteristics that make text clustering more complex. It is a highly inflected language, meaning that nouns, verbs, and

Hoti, M. H., Rexhepi, A., Hoti, A. H., Rexha, B. (2025). Challenges and Solutions in Clustering Low-Resource Language Social Media Text: An Evaluation Using Unsupervised Algorithms. *International Journal of Interactive Mobile Technologies (iJIM)*, 19(20), pp. 151–167. https://doi.org/10.3991/ijim.v19i20.56307

Article submitted 2025-04-29. Revision uploaded 2025-07-29. Final acceptance 2025-08-03.

adjectives can appear in multiple morphological forms. This increases vocabulary sparsity and makes term normalization more challenging. Additionally, regional dialects (e.g., Gheg and Tosk), code-switching with English, and a flexible word order all contribute to semantic ambiguity and non-standard expression in social media text. These features complicate both preprocessing and the ability of clustering algorithms to detect coherent topic structures without advanced linguistic modeling.

Proper and useful data clustering, especially for text data, is accompanied by several difficulties in terms of the selection of preprocessing techniques, numerical modeling, and identification of appropriate algorithms. According to [1], this is related to a methodology for processing substantial amounts of data generated by numerous entities on an ongoing basis.

This study addresses these gaps by evaluating three widely used unsupervised clustering algorithms—K-Means, DBSCAN, and HDBSCAN—on Albanian social media text data. Specifically, this study aims to explore how well these algorithms perform when applied to a language with limited resources and minimal existing NLP [2] infrastructure. Several studies, such as [3] and [4], have used and declared the most appropriate methods of classifying their data. However, this only demonstrates that other algorithms may be more appropriate for the specific experiments used in this case and does not necessarily indicate that the effectiveness of other algorithms is inferior [5], [6], and [7].

The novelty of this work lies in its focus on evaluating the suitability of traditional clustering methods in a low-resource language setting, providing insights into their effectiveness and limitations. Unlike previous studies that apply these algorithms in well-resourced language environments, this study highlights the distinct preprocessing, tokenization, and vectorization challenges encountered when working with Albanian text data. Furthermore, we examine the specific linguistic characteristics that influence clustering performance, offering a new perspective on the applicability of these methods in the broader context of low-resource NLP.

The paper is structured as follows: Section 2 outlines the related work, which is focused on the three mentioned unsupervised algorithms: K-Means, DBSCAN, and HDBSCAN. Section 3 presents the methodology and steps used in developing our testing. In Section 4, we present the results obtained, followed by discussions. Finally, Section 5 contains the conclusions of the study.

2 RELATED WORK

Classifying unannotated data poses a challenge, especially in languages with limited resources, due to the scarcity of literature and research. Therefore, specific unsupervised machine learning algorithms are well-suited for handling such data.

Based on the reviewed literature, it was found that K-Means has been implemented in studies [8], [9], and [10], DBSCAN in [11], [12], and [13]; and HDBSCAN in [14], [15], [16], [17], [18], and [19]; with numerous examples. Meanwhile, in [4] the authors conducted a systematic review of different techniques such as text mining, clustering, and identification of recent techniques for treating sentiment analysis. In [5], text mining techniques were applied to extract user experience insights from product reviews, which is conceptually similar to our use of clustering to identify user sentiment in service feedback. Similarly, [6] demonstrates the use of sentiment analysis and classification in Albanian-language social media, showing the feasibility of NLP tasks in low-resource settings. The paper [7] applies performance evaluation methods (e.g., precision, recall) to cybersecurity tasks, which align with our approach to algorithm comparison, albeit in a different domain.

According to [8], different unsupervised algorithms are used for data clustering in datasets containing low-resource languages, such as Spectral, Agglomerative, Mean Shift, and Affinity Propagation. The authors tried to identify the most effective and efficient algorithms for clustering data for two different datasets, and all algorithms performed similarly, resulting in close numbers of clusters based on the content of the datasets.

In the paper [9], unsupervised algorithms were utilized to assess various methods of automatic text document clustering for plagiarism detection. The study focused on implementing K-Means with different preprocessing techniques such as N-Grams, VSC, stemming, lemma, and chunking. Three sets of documents were used as input, taken from [10], and each combination of methods was evaluated for recall, precision, and execution time.

According to [11], it is recommended to transform unstructured data into structured data, as this can lead to higher accuracy and is better suited for subsequent analysis. In addition, the study focuses on various text representation techniques, which demonstrates that they have a direct influence on the generation of appropriate clustering algorithms. The authors of this study examined four methods of text preprocessing, including Bag of Words, TF-IDF, Word2Vec, and GloVe. The experimental results demonstrate that the effectiveness of text clustering largely depends on the text preprocessing technique used.

Ziruo Jia and Fuqiang Qi [12] utilized five types of datasets with diverse contents (such as Karate, Lesmis, Polbook, Netscience, and Metabolic). The authors also applied DBSCAN, which utilizes fast detection of centroid nodes. They measured the local density where each centroid node found corresponding nodes faster than usual, facilitating a more efficient division of other cluster nodes in the network with their corresponding nearest neighbors. Such detection, especially for central nodes, has demonstrated high effectiveness and efficiency in this application.

A modified implementation of DBSCAN is introduced in the paper [13]. The modification is based on three different layers, including 1) DBSCAN, 2) granular computing (GrC), and 3) fuzzy rule-based methods. Initially, the DBSCAN algorithm clusters data within the existing data space, and then granular processing (GrC) is employed to explain each group using information granules (IG) and reconstruction errors for rule information. The results obtained from typical datasets confirmed the effectiveness of the proposed method for such applications. Using this method can address practical challenges, especially in online recognition scenarios.

An example of the application of supervised algorithms is shown in paper [14], where the authors suggest using the Harmony Search (HS) algorithm with the SVM algorithm to improve diversity in the PSO process. HS is advantageous because it offers more diverse solutions than other methods by considering all solutions stored in memory when generating a new solution.

3 METHODOLOGY

The methodology used in this paper is based on the reviewed literature's findings and on the application and evaluation of the methods on input datasets collected from two distinct local companies. To prepare more suitable datasets, various data preprocessing methods are applied to the collected data from the official social media pages of the selected companies. However, the personal data of the authors who posted their opinions are excluded from the analysis.

We collected posts from Facebook, including user opinions and official public Facebook pages of Vala Telecommunication (https://www.facebook.com/valamobile) and Art Motion (https://www.facebook.com/artmotion.net). These two companies, based in the Republic of Kosovo, provide telecommunications and television broadcasting services in the Albanian language. Throughout the rest of this paper, we refer to these companies as X (Vala) and Y (Art Motion), respectively.

Given that the content of the posts was initially unlabeled and varied in nature, we opted to use unsupervised clustering algorithms. Based on a review of the literature and prior studies, we selected K-Means, DBSCAN, and HDBSCAN as the most relevant and commonly applied algorithms in similar settings.

We acknowledge that limiting the dataset to posts from two telecom companies may introduce domain-specific patterns. However, given the scarcity of labeled or large-scale Albanian text corpora, this dataset provided a consistent and publicly accessible source of natural language content suitable for evaluating clustering performance in low-resource settings. Future studies should explore cross-domain datasets to test the generalizability of clustering techniques in broader contexts.

Below is a Figure 1 that illustrates the steps and features employed during the preprocessing of the dataset. These steps were conducted before the application of the selected algorithms. Both datasets underwent each of these steps before clustering was applied.

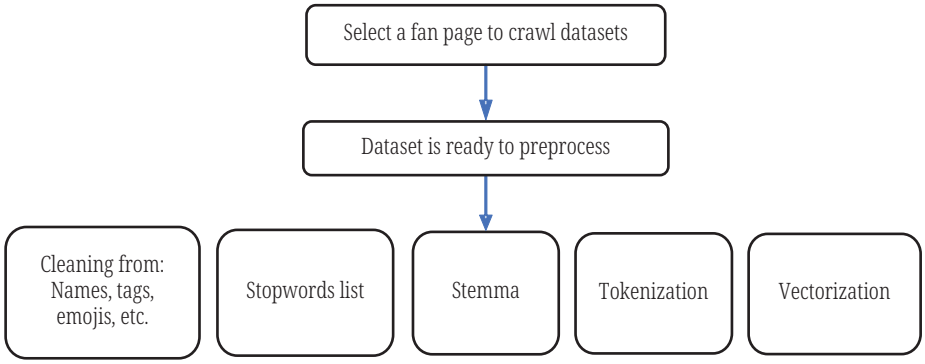

Fig. 1. Preprocessing steps of dataset

The fundamental goal of this study is to assess the performance of selected algorithms when applied to a low-resource language, such as Albanian. In particular, we aim to examine their interactions and effectiveness in processing and analyzing Albanian language data. As a result, we manually analyzed the two datasets to identify the discussed topics to assess the effectiveness of the algorithms in clustering them into separate groups.

After analyzing dataset X, we identified several topics discussed in the comments, which are represented in both English and Albanian (the original language): Vala, offer (alb: oferta), internet, problems (alb: probleme), services (alb: shërbime), validity of services (alb: validiteti i shërbimeve), package value (alb: vlera e pakove), internet costs (alb: shpenzimet e internetit), top-ups (alb: mbushjet), 3G, 4G, fees (alb: tarifat), praise for customer service (alb: lëvdata për shërbime të konsumatorëve), and network (alb: rrjeti).

For dataset Y, the following topics were identified: ArtMotion, network (alb: rrjeti), channels (alb: kanalet), validity of offers (alb: vlefshmëria e ofertave), quality of services (alb: kualiteti i shërbimeve), offers (alb: ofertat), services (alb: shërbimet), price increase (alb: shtrenjtimi), broadcasting (alb: transmetimet), prices (alb: çmimet), internet, commentator (alb: komentator), etc.

In order to increase accuracy and achieve better results in the preprocessing stage, we performed various data cleaning operations, such as cleaning tags, names, emojis, and so on. We also manually reviewed and normalized common slang and spelling variants (e.g., afati → afat) and removed or flagged code-switched terms (English/Albanian mix) that were not frequent enough to be retained in the final vocabulary. Next, we converted the data into numerical values and vectors using TF-IDF and Word2Vec.

Meanwhile, we chose to use TF-IDF for the vectorization of the Albanian text data. While modern embeddings like BERT and GPT have shown superior performance in various NLP tasks, they are not always the best choice for low-resource languages such as Albanian. One of the main reasons for opting for TF-IDF is the limited availability of pre-trained models specifically for the Albanian language. Although some multilingual models exist, they are often not fine-tuned for less common languages, which can result in suboptimal performance.

Furthermore, TF-IDF remains a straightforward and effective method for text representation, especially when working with smaller datasets and simpler clustering algorithms. It allowed us to maintain control over the vocabulary and ensure that relevant features from the Albanian language were captured accurately, without the risk of relying on embeddings that may not be well-suited for this specific language.

In future work, we plan to experiment with more advanced embeddings like BERT, but for this study, TF-IDF provided a solid baseline for evaluating clustering performance. Although transformer-based models like multilingual BERT (mBERT) and XLM-R have achieved impressive results in high-resource NLP tasks, their application to low-resource languages such as Albanian remains challenging without domain-specific fine-tuning. Preliminary experiments using mBERT embeddings resulted in less coherent clusters, likely due to insufficient language-specific training and noise sensitivity. Recent studies [15], [16] have also reported that off-the-shelf multilingual embeddings often underperform in morphologically complex, low-resource settings.

In contrast, TF-IDF allowed us to maintain control over feature selection, capture meaningful term-frequency relationships, and generate consistent clustering results in our targeted domain.

For the evaluation of the clustering algorithms, we used Silhouette scores as the primary quantitative metric. In addition, we conducted a manual evaluation by analyzing the most frequent terms and sample posts within each cluster to verify their thematic coherence. While we refer to this as 'accuracy' in some parts of the text, we clarify that this reflects a qualitative, post hoc assessment of cluster interpretability, not a classification-based accuracy score.

However, we also used the Silhouette score to capture the quality of the clusters, which evaluates how similar each point is to its own cluster compared to other clusters. This combination of metrics allowed us to assess both the overall performance and the internal consistency of the clusters.

The use of these metrics is critical in understanding the practical application of clustering in real-world settings. For example, in customer feedback analysis for telecom companies such as X and Y, accuracy helps determine whether customer complaints about similar issues are being grouped together. The Silhouette score, on the other hand, helps ensure that the clusters are distinct and well-formed, which is important for identifying clear trends or issues within the feedback.

Additionally, we compared these metrics across three different algorithms (K-Means, DBSCAN, and HDBSCAN) to provide a comprehensive evaluation framework. This comparative analysis highlights the trade-offs between the algorithms,

such as K-Means excelling in overall accuracy but struggling with outliers, whereas DBSCAN and HDBSCAN showed better performance in handling noisy data. These insights are crucial for selecting the right clustering approach in practical applications where clean data is not always guaranteed.

4 RESULTS AND DISCUSSIONS

While the algorithms tested in this study—K-Means, DBSCAN, and HDBSCAN—are well-established in the field of clustering, their application to low-resource languages such as Albanian offers valuable insights. One of the main challenges with these algorithms is their reliance on high-quality, well-preprocessed data, which is often lacking in low-resource language scenarios. The absence of large, annotated datasets and pre-trained models makes it difficult for these algorithms to achieve optimal results.

For instance, K-Means showed strong performance in terms of accuracy, but its sensitivity to outliers and inability to handle varying cluster densities limit its effectiveness in noisy, real-world datasets common in low-resource languages. DBSCAN and HDBSCAN, while better suited for handling outliers and dense clusters, still struggle with the sparse and uneven distribution of data points typical in these settings.

To improve the performance of clustering algorithms for low-resource languages, future work could focus on customizing these methods to account for the specific challenges of such languages. For example, developing hybrid algorithms that combine the density-based strengths of DBSCAN with the computational efficiency of K-Means could yield better results. Additionally, incorporating domain-specific knowledge or fine-tuning these algorithms for low-resource languages through techniques like transfer learning could further enhance clustering accuracy and robustness.

Ultimately, while the algorithms used in this study are not new, applying them to low-resource languages reveals important limitations that require further exploration. Addressing these issues could lead to more accurate clustering methods that work better in low-data environments, contributing to the overall improvement of NLP [17] for lesser-studied languages.

Each algorithm tested—K-Means, DBSCAN, and HDBSCAN—showed different strengths and weaknesses when applied to our dataset. K-Means performed the best in terms of overall accuracy, achieving higher clustering scores in both datasets. This is largely because K-Means is efficient at forming well-defined clusters when the data is relatively clean. However, K-Means struggled with outliers, often placing noisy data points into incorrect clusters or forcing them into existing groups, which negatively impacted the clustering quality.

DBSCAN, on the other hand, was highly effective in isolating outliers, thanks to its density-based approach. Unlike K-Means, which assigns all points to clusters, DBSCAN can leave noisy points unassigned, reducing distortion of true cluster structure. However, this strength comes at the cost of sensitivity to eps and minPts parameters, which can lead to fragmented or incoherent clusters if not carefully tuned—especially in sparse text data like ours.

HDBSCAN offered a balance between the two, providing better handling of varying densities in the data. Unlike K-Means, it does not require a set number of clusters, and unlike DBSCAN, it adjusts the cluster shape based on the data's density. This made it more flexible when dealing with both sparse and dense regions in

our datasets. However, HDBSCAN was slower than the other two algorithms, making it less suitable for real-time applications or larger datasets.

It is important to mention that during manual inspection, we observed several patterns in misclustered data. Clustering errors often appeared in short, vague, or context-dependent comments, such as those using sarcasm, idioms, or slang. For example, user posts containing words like "service" or "price" were sometimes placed into incorrect clusters because those terms were common across both positive and negative feedback. Additionally, overlapping vocabulary—such as the term "network," which could refer to internet, mobile signal, or TV services—led to topic blending in some clusters. These observations highlight the challenge of clustering informal, domain-specific social media text in a low-resource language. Future work could mitigate such issues by incorporating linguistic features, contextual embeddings, or sentiment-aware models to refine cluster separation.

4.1 K-Means clustering algorithm

During the implementation of the K-Means algorithm, we encountered several inconvenient clusters (outliers) and instances of confusion. We identified that these outliers and inappropriate clusters stemmed from word variations, such as "afati," "afatin," "afatet," and "afatit." Consequently, we decided to replace these words with their stem, "afat," in order to mitigate this issue.

This transformation was beneficial, resulting in improved accuracy and clustering. We utilized the estimation values of minimum and maximum, referred to as "corpus-specific stop words," to extract data effectively and achieve optimal results. In specific instances, the minimum value criterion was set to ignore words that appeared in fewer than five comments, with the appropriate value for this criterion set to min = 5. Similarly, the maximum value criterion was set to disregard phrases that appeared in more than 50% of comments, with the appropriate value set to max = 0.5.

Following these actions, we obtained results of 3 to 5 clusters, which we deemed the optimal number of clusters. To confirm this conclusion, we applied the Elbow and Silhouette methods to verify the number of clusters. Both methods indicated that the obtained cluster numbers were optimal for the content used.

For dataset X, the comments (Facebook posts) are classified into five clusters, numbered 0 to 4. In the case of dataset X, most of the posts (compliments and complaints) revolve around concepts such as Vala (name of the company), services (alb: shërbimet), Internet, and prices (alb: çmimet). Dataset X contains 1,326 posts, with the majority grouped into Cluster 2 (discussed topics are mainly related to the internet). This result initially seemed biased; therefore, upon further analysis of the posts, we found a real correlation between the posts in that cluster. Figure 2 visualizes these results. For the second dataset, Y, the results are as follows: the posts are grouped into three clusters—Cluster 0 contains 365 posts, followed by Cluster 1 with 39 posts, and Cluster 2 with 21 posts. A similar dilemma, as in the first dataset, arose with Cluster 0 in the second dataset, Y. However, this was resolved by analyzing the posts in that cluster. We believe these biases are typical due to people's habits, since, usually, when they discover a problem or topic, it becomes the focus, and most discussions and writing are directed towards it. The most discussed topics identified in dataset Y include Art Motion, channels, and network (alb: kanale dhe rrjet), and the provision of services (alb: ofrimi i shërbimeve).

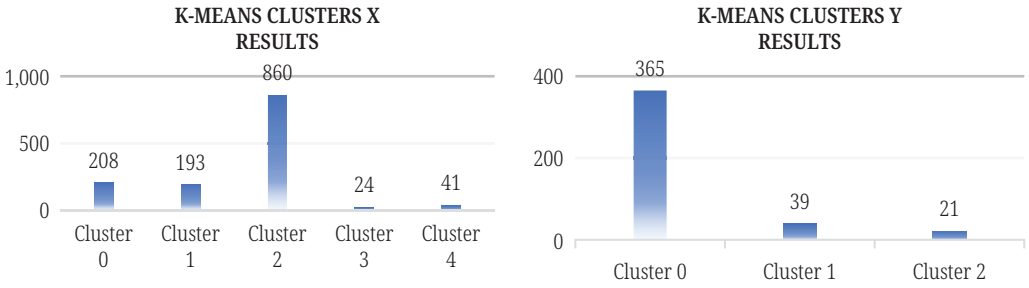

Fig. 2. K-Means algorithm clustering results for dataset X (left) and dataset Y (right)

To ensure the posts were correctly clustered and the results were acceptable, based on research [18], we employed the Elbow and Silhouette methods to verify the number of clusters and the results obtained. The Elbow and Silhouette check tests were also reported in several other studies, such as [19], [20], and more.

Using the Elbow method, we identified optimal cluster numbers between 3 and 5 for both datasets. To assess cluster quality, we performed a manual post hoc evaluation based on topic coherence, which yielded consistent thematic groupings. These were estimated to correspond to approximately 70.1% and 76.4% interpretability for datasets X and Y, respectively. The next check applied the Silhouette method, and the outcomes are displayed in Table 1.

In Table 1, we see that for dataset X, the most suitable number of clusters is 5 (Silhouette score = 0.701). For dataset Y, the optimal number is 3 (Silhouette score = 0.764). These results align with the thematic patterns observed during manual cluster interpretation.

The Silhouette scores across K-Means, DBSCAN, and HDBSCAN indicate different clustering behaviors. K-Means achieved the highest Silhouette scores for both datasets (0.701 for X and 0.764 for Y), reflecting its strength in generating compact clusters under clean, vectorized inputs. DBSCAN showed lower but reasonable scores, particularly for dataset Y (0.688), where it handled noise more effectively. HDBSCAN results were generally close to DBSCAN but produced a larger number of small or outlier clusters, leading to slightly reduced average scores. Although we did not compute confidence intervals, the consistent trends across both datasets support the observed algorithmic trade-offs. This suggests that the K-Means method is effective for a low-resource language such as Albanian. Figure 3 shows the Elbow graphs, and the Silhouette graph is shown in Figure 4.

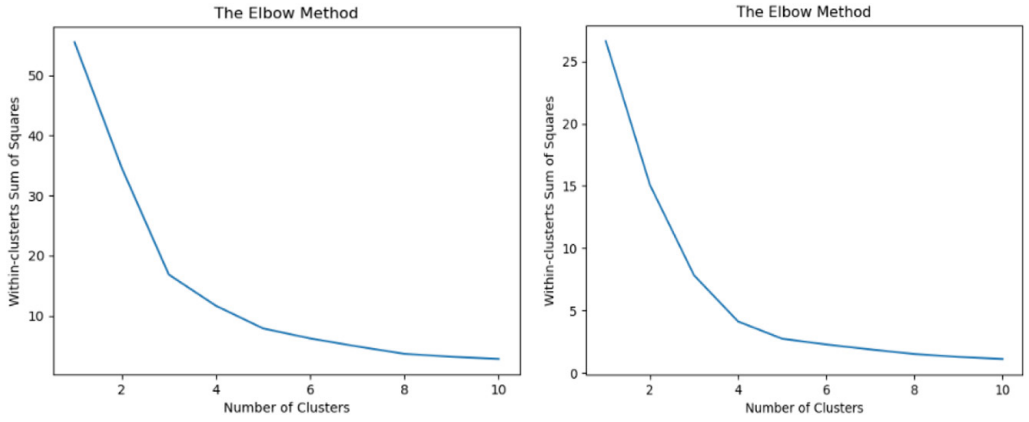

Fig. 3. K-value selection algorithms of clustering with the Elbow method for dataset X (left) and dataset Y (right)

Table 1. Silhouette accuracy score for X and Y datasets (left side X, right side Y)

Silhouette Clusters Accuracy of K-Means		
No. of Clusters	**X**	**Y**
2	0.599	0.716
3	0.680	0.764
4	0.691	0.734
5	0.701	0.741
6	0.688	0.740
7	0.677	0.678
8	0.690	0.680
9	0.682	0.461
10	0.635	0.500

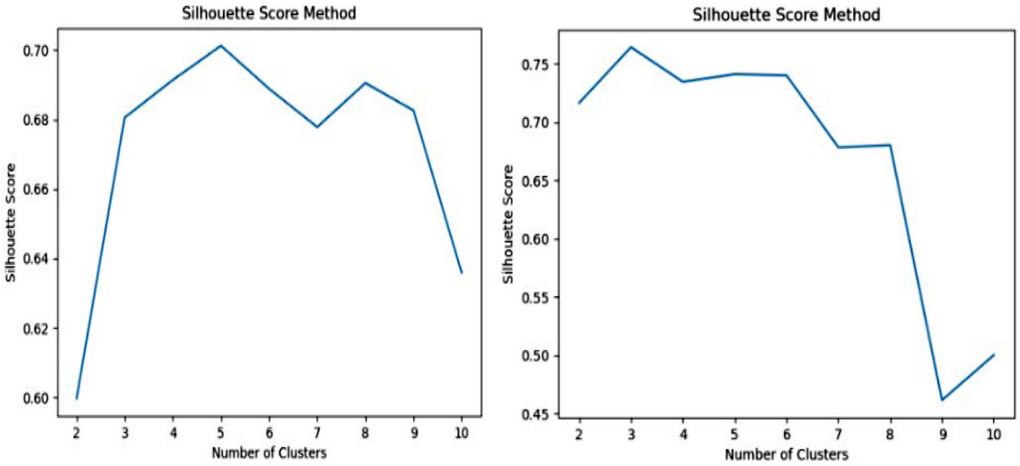

Fig. 4. Silhouette graphs for dataset X (left) and dataset Y (right)

A visual representation of the results of both datasets is shown in Figure 5, with colors identifying the clusters.

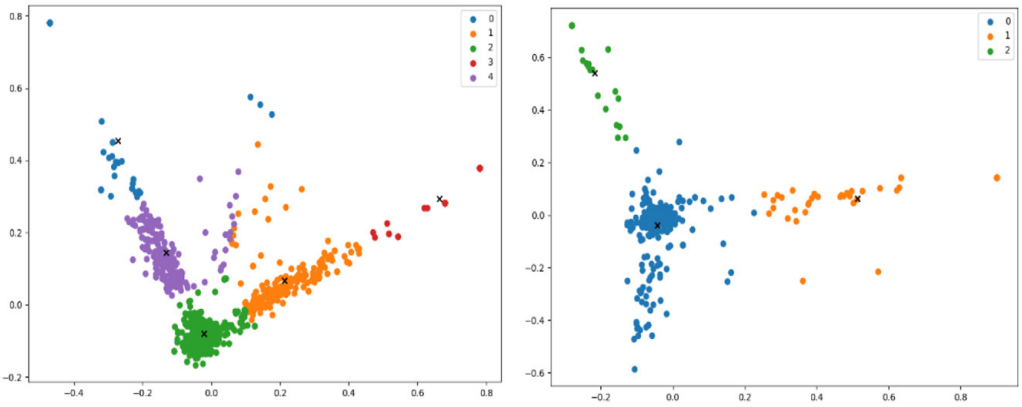

Fig. 5. Visualization of the results for dataset X (left) and dataset Y (right)

4.2 DBSCAN clustering algorithm

DBSCAN is a popular method for clustering complex and multifaceted items. It is used for a broad range of applications and different issues, including ultra-classification [21], three-dimensional point cloud commercial airliner categorization [22], and so on.

DBSCAN calculates each value, as it is one of the algorithms focusing on the density of data and concentration in areas with the highest discussion. This is achieved based on assessment rather than initial parameters for group predictions. Based on our testing results using both datasets, we propose a quick and accurate implementation of DBSCAN for low-resource languages, along with the corresponding application steps. DBSCAN relies on two parameters, eps and minPoints, which need to be adjusted to generate satisfactory results.

The parameter "eps" defines the radius of neighbors around a point, while minPoints specifies the minimum number of neighbors within the "eps" radius. After careful preprocessing of the collected dataset, a list of stopwords was applied, followed by tokenization. Additionally, all words were transformed from uppercase to lowercase and divided into N-Grams before being vectorized. We applied PCA to reduce the dimensionality of the TF-IDF matrix to two principal components (P1 and P2) for visualization and to better understand the structure of the data. The total explained variance provided insight into how well the reduced data preserved the original variance. While PCA was not used to mathematically compute eps, visual inspection of the PCA scatterplots allowed us to estimate reasonable eps values for clustering, based on inter-point distances in the 2D projection. This was a heuristic approach used to guide parameter tuning in the absence of labeled data or standardized tuning procedures:

X dataset – Total sum of variance of P1 and P2: [0.023741 0.04330879]
Y dataset – Total sum of variance of P1 and P2 [0.04171457 0.07738584]

As can be seen, for the X dataset, the appropriate eps value is 0.23741, while for the Y dataset, it is 0.076715. These results are presented in Table 2, where all the parameters used to identify the most appropriate eps value are shown. In Figure 6, the Elbow graph was used to determine the optimal cluster significance. This is indicated by the point at which the graph bends the most, representing the ideal number of clusters.

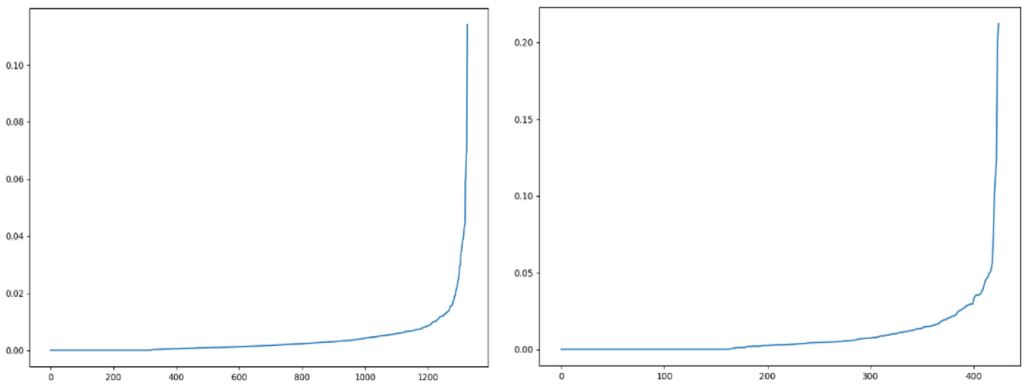

Fig. 6. Elbow graph for identifying the appropriate number of clusters for dataset X (left) and dataset Y (right)

For dataset X, a total of eight clusters were generated, but 74 comments were not assigned to any of them and were classified as outliers. However, the interpretability of DBSCAN's clusters—based on manual topic evaluation—was lower, estimated at around 50%, particularly for dataset X. This was consistent with the lower Silhouette score, reflecting challenges in forming coherent clusters in sparse text data. The topics discussed in dataset X include internet, services (alb: shërbime), Vala, offers (alb: oferta), the validity of the packages, prices (alb: çmimet), network (alb: rrjeti), 3G, 4G, etc. In the case of dataset Y, six clusters were generated, with 358 comments in cluster 0, 31 in cluster 1, 9 in cluster 5, and a few in the others.

The topics discussed in this dataset include offers (alb: oferta), prices (alb: çmimet), services (alb: shërbimet), commentary (in the sense of game commentator—alb: komentim), Art Motion, services validity (alb: vlefshmëria e shërbimeve), and services quality (alb: kualiteti i shërbimeve). The outcomes are displayed in Figure 7, and the accuracy expressed as a percentage is 68.8%, which is higher than the accuracy in the case of dataset X.

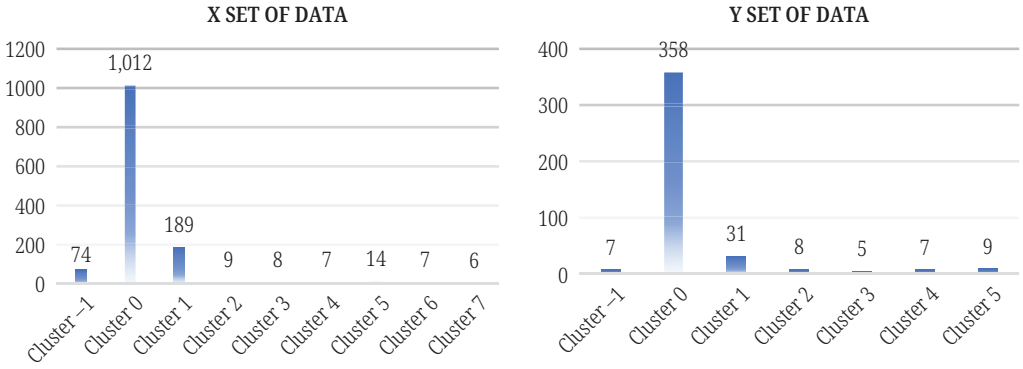

Fig. 7. DBSCAN clustering using dataset X (left) and dataset Y (right)

To identify the epsilon value and minimum samples for the DBSCAN algorithm, we used the P1 and P2 components to indicate the points in space that were used to generate this clustering form (see Figure 8).

From the obtained variance interval, we aimed to find the most appropriate values for eps and minPoints. Table 2 shows an attempt to identify these values for both datasets. The third row gives these values for the X dataset, while in the seventh row, we find the values for the Y dataset.

Table 2. Silhouette accuracy score for dataset X (left) and dataset Y (right)

Prediction of Suitable PCA Using the Appropriate Number of Clusters for DBSCAN									
X Set of Data					Y Set of Data				
Test Case	No. of Clusters	Silhouette Score	Epsilon Value	Minimum Points	Test Case	No. of Clusters	Silhouette Score	Epsilon Value	Minimum Points
0	22	0.1334	0.023741	2	0	17	0.5532331	0.042715	2
1	15	0.1289	0.023741	3	1	12	0.538805	0.042715	3
2	11	0.3532	0.023741	4	2	12	0.532972	0.042715	4
3	9	**0.4862**	**0.023741**	5	3	11	0.513775	0.042715	5
4	22	0.1338	0.024741	2	4	17	0.554859	0.042715	2

(Continued)

Table 2. Silhouette accuracy score for dataset X (left) and dataset Y (right) *(Continued)*

Prediction of Suitable PCA Using the Appropriate Number of Clusters for DBSCAN									
X Set of Data					Y Set of Data				
Test Case	No. of Clusters	Silhouette Score	Epsilon Value	Minimum Points	Test Case	No. of Clusters	Silhouette Score	Epsilon Value	Minimum Points
...	...	...	...	...	...	...	...	...	...
71	8	0.3156	0.040741	5	71	**7**	**0.688035**	**0.076715**	**5**
72	14	0.2631	0.041741	2	72	8	0.61088	0.076715	2
73	10	0.2623	0.041741	3	73	7	0.688035	0.076715	3
74	10	0.2624	0.041741	4	74	7	0.688035	0.076715	4
75	8	0.3156	0.041741	5	75	7	0. 688035	0.076715	5

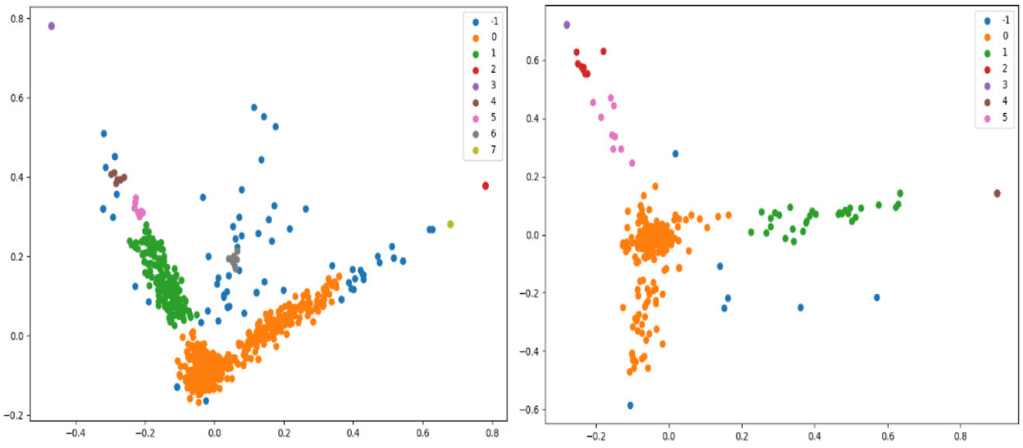

Fig. 8. Visualization of the results for dataset X (left) and dataset Y (right)

4.3 HDBSCAN clustering algorithm

The HDBSCAN algorithm is an extension of DBSCAN that builds a hierarchical representation of clusters based on varying densities. Unlike DBSCAN, which uses a single density threshold, HDBSCAN is capable of identifying clusters of differing density levels and extracting the most stable structures. This makes it particularly well-suited for sparse and noisy datasets such as social media text.

Using the epsilon value and minimum sample parameters, we identified the most frequent terms in the X dataset, which include “Vala,” “internet,” “3G,” “4G,” “services,” and others. These terms were used as centroids for the clustering process. On the other hand, for dataset Y, the results include services that are not offered despite clients paying for them, Art Motion services, validity of services, price, etc.

The HDBSCAN algorithm places a strong emphasis on identifying clusters of data that exhibit commonalities in the underlying patterns they contain. To achieve this goal effectively, the variance ratio of the entire dataset was calculated. The value obtained is:

The total sum of the variance of X dataset for P1 and P2 [0.02374082 0.04330884]

The total sum of the variance of Y dataset for P1 and P2 [0.04171457 0.07738584]

Additionally, the P1 and P2 components are within the same range as those obtained from the ideal PCA.

The results of the comment clusters identified by their content showed differences compared to the other algorithms selected for this study. These differences were mainly related to the number of clusters generated by HDBSCAN, which included clusters with a small number of comments known as outliers. This reduced the accuracy and effectiveness of the algorithm compared to the other algorithms assessed and analyzed. In both cases, the most suitable classification [23] was found to be 5–6 clusters.

In our case, we obtained the same number of clusters as discussed above, which is the ideal number of clusters as in the other selected algorithms. The clustering results for both datasets are shown in Figures 9 and 10.

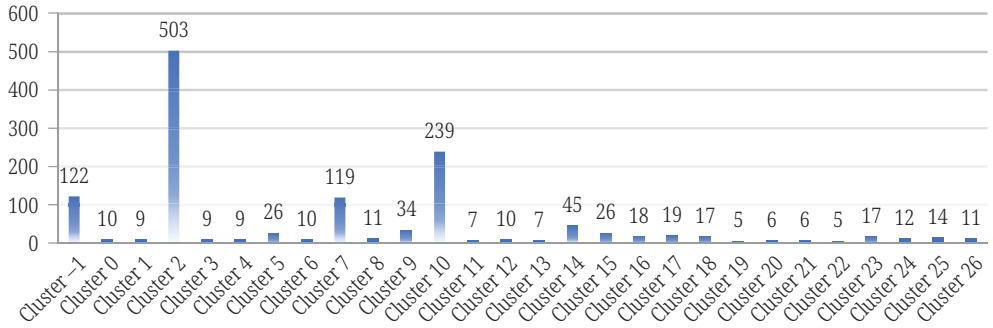

Fig. 9. HDBSCAN clustering for dataset X

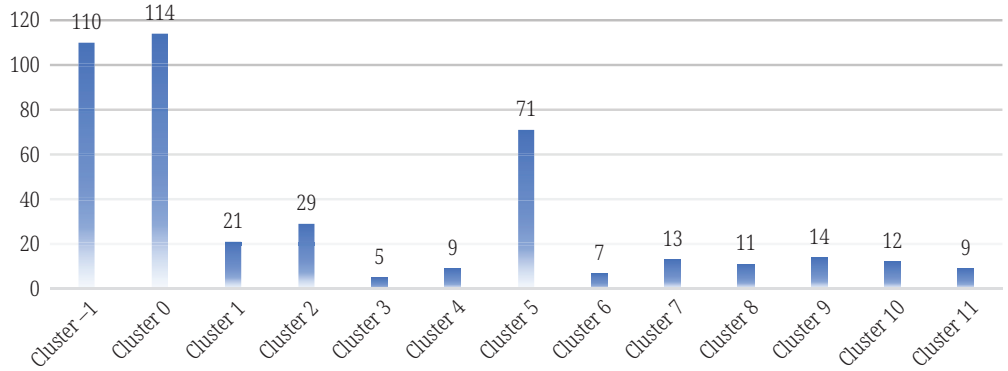

Fig. 10 HDBSCAN clustering for dataset Y

The obtained results are also visualized in Figure 11, making the clusters clearer and more visible. This means that the clusters with the highest number of comments are those with the most similar content related to the discussed topic, resulting in accurate clustering of the datasets.

In both cases, distinct colors are assigned to all clusters, representing the potential groups named in the left part of the visualization. The clustering visualizations for both datasets are shown in Figure 11.

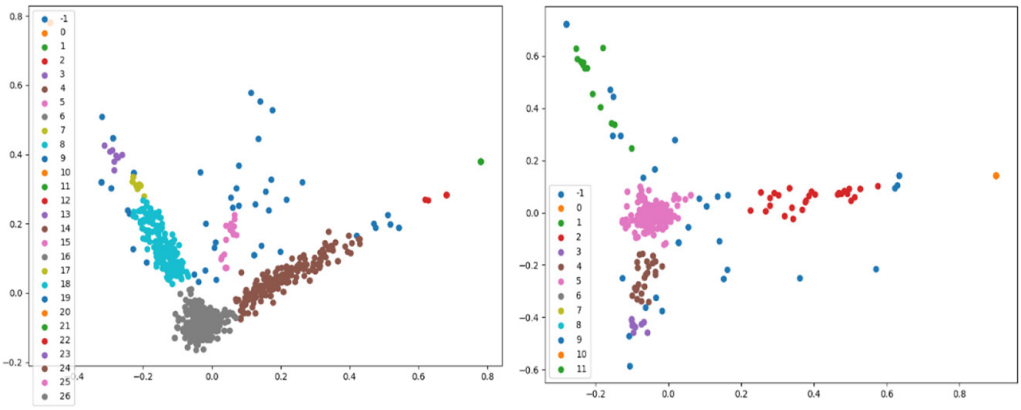

Fig. 11. Visualization form of HDBSCAN clustering of dataset X (left) and dataset Y (right)

5 CONCLUSIONS

The findings of this study have practical implications, especially for businesses and organizations working in Albanian-speaking regions. For instance, the clustering results can be directly applied to sentiment analysis for local companies like Vala and Art Motion. By identifying customer opinions on various services, these companies can better understand customer satisfaction and identify key areas for improvement, such as network quality or pricing strategies. The use of clustering helps break down large volumes of social media feedback into actionable categories, making it easier for businesses to focus on specific problems.

To qualitatively assess the coherence of the generated clusters, we manually reviewed representative posts and top terms within each cluster. This allowed us to evaluate whether each cluster reflected a meaningful topic or theme relevant to the telecom services context. While this was not a formal annotation process, it provided a practical measure of interpretability in the absence of ground-truth labels. Furthermore, to validate the number of clusters, we utilized the Elbow and Silhouette methods.

Initially, we employed the K-Means algorithm, as a more popular unsupervised algorithm, on our preprocessed datasets X and Y. This yielded 3 to 5 clusters, which were deemed optimal clustering numbers. While analyzing the results, we encountered two challenges: the first was the number of clusters identified, which was confirmed by using the Elbow and Silhouette methods and additionally by manual analysis, whereas the second was biased clusters. For the second issue, we think that biases are a result of human tendencies.

For example, sometimes when someone writes a topic, others are inclined to comment on it. For both our datasets, the number of identified clusters fell within the range of the numbers found, and the accuracy was 70.1% and 76.4% for datasets X and Y, respectively. For the second dilemma, we manually examined the biased cluster and found that all posts were thematically correlated and relevant to the same topic. Subsequently, we conducted a second experiment using the DBSCAN algorithm on the same datasets, X and Y.

The interpretability of DBSCAN's clusters in dataset X was relatively lower (~50%), largely due to data sparsity and sensitivity to parameter tuning. This lower clustering quality does not reflect a flaw in the algorithm but rather highlights the challenge of applying density-based methods in sparse text environments. For dataset Y, six clusters were identified, with an improved accuracy of 68.8%.

Our third experiment involved using the HDBSCAN algorithm, through which we identified parameters that enhance the level of similarity among comments

within each cluster. This study provides a comparative analysis of the performance of K-Means, DBSCAN, and HDBSCAN on low-resource language text data, specifically focusing on social media posts in Albanian. Our findings show that while K-Means consistently performs well in terms of accuracy, DBSCAN and HDBSCAN offer better handling of outliers and data density, respectively. Importantly, the distinct behavior of these algorithms in the context of a low-resource language like Albanian highlights the necessity of further refinement and customization for this specific use case.

The contribution of this study extends beyond a simple evaluation of clustering algorithms; it offers valuable insights into the unique challenges faced when working with low-resource languages. As more NLP tools and models are developed for less commonly studied languages, this study serves as a foundation for understanding how unsupervised clustering can be effectively applied, as well as how existing algorithms can be adapted for better performance in low-resource environments.

One limitation of this study is the use of domain-specific data, which may constrain generalization across other sectors or text types. Future work can expand this study in several directions. First, the integration of more recent techniques such as BERTopic and transformer-based embeddings (e.g., multilingual BERT, XLM-Roberta, or LASER) may enhance semantic clustering in noisy or unstructured text. However, due to the limited availability of high-quality, fine-tuned embeddings for the Albanian language, our current study prioritized traditional algorithms that offer better interpretability and control in low-resource NLP environments.

Second, future research could involve developing domain-adapted embeddings specifically fine-tuned on Albanian social media text to capture linguistic nuances and colloquialisms more effectively.

Third, embedding comparisons between TF-IDF and contextualized models should be conducted to evaluate performance trade-offs in low-resource contexts.

Fourth, the use of external clustering evaluation metrics such as Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and purity could provide a more rigorous assessment—particularly when labeled datasets or domain taxonomies are available. Additional internal evaluation metrics, such as the Davies-Bouldin Index or Calinski-Harabasz score, may be used in future work to complement the Silhouette score and offer more robust assessments of clustering performance. Additionally, semi-supervised or weakly supervised clustering methods could be explored to mitigate the challenge of label scarcity in low-resource settings.

Finally, future work should expand the study to include datasets from multiple domains—such as health, education, and e-government services—to assess the generalizability and robustness of clustering algorithms across different application areas. These directions will contribute to building more adaptable and effective clustering systems for underrepresented languages.

6 REFERENCES

[1] S. Abdulah, W. Atwa, and A. M. Abdelmoniem, "Active clustering data streams with affinity propagation," *The Korean Institute of Communications and Information Sciences (KICS),* vol. 8, no. 2, pp. 276–282, 2022. https://doi.org/10.1016/j.icte.2021.08.017

[2] N. El Rhezzali, I. Hilal, and M. Hnida, "NLP-enhanced techniques for cheating detection in virtual exams: A comparative study of string and semantic similarity measures with K-Shingling, Minhashing, LSH, and K-Means," *International Journal of Interactive Mobile Technologies (iJIM),* vol. 19, no. 3, pp. 56–72, 2025. https://doi.org/10.3991/ijim.v19i03.49897

[3] D. Abdullah, S. Susilo, A. S. Ahmar, R. Rusli, and R. Hidayat, "The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data," *Quality & Quantity,* vol. 56, pp. 1283–1291, 2021. https://doi.org/10.1007/s11135-021-01176-w

[4] M. H. Hoti *et al.*, "Text mining, clustering and sentiment analysis: A systematic literature review," in *2022 11th Mediterranean Conference on Embedded Computing (MECO)*, Budva, Montenegro, 2022.

[5] K.-Y. Lin, "A text mining approach to capture user experience for new product development," *International Journal of Industrial Engineering: Theory, Applications and Practice,* vol. 25, no. 1, p. 108, 2018. https://doi.org/10.23055/ijietap.2018.25.1.4014

[6] M. H. Hoti, H. Hoti, and E. Kurhasku, "Sentiment analysis of positive and negative comments, extracted from social networks and web in Albanian language," *International Journal of Applied Systemic Studies,* vol. 11, no. 2, pp. 1–15, 2024. https://doi.org/10.1504/IJASS.2024.140018

[7] V. Bytyqi and B. Rexha, "Machine learning boosted trees algorithms in cybersecurity: A comprehensive review," in *Advances in Information and Communication. FICC 2024.* in Lecture Notes in Networks and Systems, K. Arai, Ed., vol. 920, Springer, Cham, 2024, pp. 158–173. https://doi.org/10.1007/978-3-031-53963-3_12

[8] M. H. Hoti *et al.*, "Spectral analysis, agglomerative, mean shift and affinity propagation algorithms, use on the content from social media for low-resource languages," in *2023 46th MIPRO ICT and Electronics Convention (MIPRO)*, Opatija, Croatia, 2023.

[9] K. Vani and D. Gupta, "Using K-means cluster based techniques in external plagiarism detection," in *2014 International Conference on Contemporary Computing and Informatics (IC3I),* Mysore, India, 2014, pp. 1268–1273. https://doi.org/10.1109/IC3I.2014.7019659

[10] M. Potthast *et al.*, "Overview of the 5th international competition on Plagiarism detection," in *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers,* Valencia, Spain, 2013.

[11] K. S. Aswhini, C. P. Shantala, and T. Jan, "Impact of text representation techniques on clustering models," *Research Square,* 2022. https://doi.org/10.21203/rs.3.rs-1385057/v1

[12] Z. Jia and F. Qi, "Network clustering algorithm based on fast detection of central node," *Scientific Programming,* vol. 2022, pp. 1–5, 2022. https://doi.org/10.1155/2022/4905190

[13] X. Zhang, X. Shen, and T. Ouyang, "Extension of DBSCAN in online clustering: An approach based on three-layer granular models," *Applied Sciences,* vol. 12, no. 19, pp. 1–21, 2022. https://doi.org/10.3390/app12199402

[14] J. Han and Y. Seo, "Feature selection and parameter optimization for support vector machines using particle swarm optimization and harmony search," *International Journal of Industrial Engineering: Theory, Applications and Practice,* vol. 28, no. 1, pp. 1–7, 2021.

[15] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is Multilingual BERT?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* Florence, Italy, 2019. https://doi.org/10.18653/v1/P19-1493

[16] S. Wu and M. Dredze, "Are all languages created equal in Multilingual BERT?" in *Proceedings of the 5th Workshop on Representation Learning for NLP,* in Online Association for Computational Linguistics, 2020.

[17] M. H. Hoti, F. Qorrolli, and F. Spahija, "Enhancing fake news detection via stance analysis: Leveraging advanced NLP techniques and machine learning models," *International Journal of Interactive Mobile Technologies (iJIM),* vol. 19, no. 11, pp. 39–50, 2025. https://doi.org/10.3991/ijim.v19i11.55007

[18] J. A. Lossio-Ventura, J. Morzan, H. Alatrista-Salas, T. Hernandez-Boussard, and J. Bian, "Clustering and topic modeling over tweets: A comparison over a health dataset," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, San Diego, CA, USA, 2019. https://doi.org/10.1109/BIBM47256.2019.8983167

[19] M. U. Cakir and S. Guldamlasioglu, "Text mining analysis in Turkish language using big data tools," in *IEEE 40th Annual Computer Software and Applications Conference,* Atlanta, USA, 2016. https://doi.org/10.1109/COMPSAC.2016.203

[20] C. Yuan and H. Yang, "Research on K-value selection method of K-means clustering algorithm," *J Multidiscplinary Scientific Journal,* vol. 2, no. 16, pp. 226–235, 2019. https://doi.org/10.3390/j2020016

[21] J. Shen, X. Hao, Z. Liang, Y. Liu, W. Wang, and L. Shao, "Real-time superpixel segmentation by DBSCAN clustering algorithm," *IEEE Transactions on Image Processing,* vol. 25, no. 12, pp. 5933–5942, 2016. https://doi.org/10.1109/TIP.2016.2616302

[22] H. Chen, M. Liang, W. Liu, W. Wang, and P. X. Liu, "An approach to boundary detection for 3D point clouds based on DBSCAN clustering," *Pattern Recognition,* vol. 124, no. C, pp. 1–8, 2022. https://doi.org/10.1016/j.patcog.2021.108431

[23] T. A. Sandy, A. Ghufron, A. Muhtadi, and Pujiriyanto, "Text classification of Duolingo reviews on Google Play: Insights for enhancing M-learning applications," *International Journal of Interactive Mobile Technologies (iJIM),* vol. 19, no. 7, pp. 206–223, 2025. https://doi.org/10.3991/ijim.v19i07.52891

7 AUTHORS

Mërgim H. Hoti is Dr. Sc. in Computer Science. He is currently a Teaching Assistant at the University of Prishtina "Hasan Prishtina," Faculty of Electrical and Computer Engineering. His research interests include Data Science, Machine Learning, Artificial Intelligence, and Cybersecurity, with several publications in these areas (E-mail: mergim.hoti@uni-pr.edu).

Avni Rexhepi is a Professor at the University of Prishtina "Hasan Prishtina," Faculty of Electrical and Computer Engineering and head of "Computer and Software Engineering" department. His field of interest encompasses Programming, Data Structures and Algorithms and AI (E-mail: avni.rexhepi@uni-pr.edu).

Arbër H. Hoti is a Teaching Assistant at the University of Prishtina "Hasan PRISHTINA," Faculty of Education. Also, he is in the end of doctoral studies at South East European University in Computer Science. His research interests' includes Machine Learning, Artificial Intelligence and LMS where he has several papers indexed in digital platforms (E-mail: arber.hoti1@uni-pr.edu).

Blerim Rexha is a Professor at the University of Prishtina, Faculty of Electrical and Computer Engineering in Prishtina, Kosovo. He boasts of an impressive publication record in respected international conferences and journals. He's also been a guest speaker at numerous national and international conferences, highlighting his academic and research experience, includes biometrics, privacy, cybersecurity, cryptography, and machine learning (E-mail: blerim.rexha@uni-pr.edu).