Spectral Analysis, Agglomerative, Mean Shift and Affinity Propagation Algorithms, Use on the Content from Social Media for Low-Resource Languages

Mërgim H. HOTI, Jaumin Ajdari, Xhemal Zenuni, Mentor Hamiti
South East European University (SEEU)/ Computer Science, Tetovo, Republic of North Macedonia
{mh28356, j.ajdari, xh.zenuni, m.hamiti}@seeu.edu.mk

***Abstract* - Social networks, as part of our daily life, affect our behavior and lifestyle in various ways, making it important for each of us to be aware of their impact. Posts on social media platforms can have a profound effect on our mood, depending on our personal interpretation and opinion of them. Therefore, it is crucial to correctly classify these textual data to gain a better understanding of their impact. However, this task can be challenging, particularly when dealing with unlabeled data such as social media posts. An added challenge is working with low-resource languages.**

In this research, we investigate four unsupervised text clustering methods by testing them on a low-resource language, such as Albanian. The investigated algorithms are Spectral, Agglomerative, Mean Shift and Affinity Propagation, and by adjusting the working parameters, we tried to find a more appropriate application of them. Methods are applied to pre-processed data, textual posts, by use of different preprocessing techniques, and the results are presented and interpreted.

This research aims to assist other researchers in the same field who have a specific focus on working with low-resource languages.

Keywords - text classification, low-resource language, Agglomerative, Spectral Analysis, Affinity Propagation, Mean Shift..

I. INTRODUCTION

Nowadays, there is a huge increase in social content, data that is distributed on many social networks such as Facebook, Instagram, Twitter, LinkedIn, Tik-Tok, and so on. Those data are very extensive and require proper tools for effective management. The proper management and processing of these data make today's life easier, especially in the process of decision-making. Knowing the opinions of customers is of great value in business. Therefore, the correct classification of data is a step in helping companies to make changes in services and customer satisfaction.

To the best of our knowledge, there is no literature available that has worked specifically with unsupervised clustering algorithms using low-resource languages, excepted to some researchers which have used Albanian as low resources language on their papers such as [1], [2], [3], therefore, we decided to test several unsupervised clustering algorithms on a low-resource language such as Albanian. The main idea of the research was to analyze the correctness and accuracy of these algorithms in the dataset of posts, on the social network (Facebook), written in the Albanian language.

After analyzing some existing literature on this issue, we decided to focus our research on unsupervised [4] clustering methods and used clustering methods such as spectral [5, 6, 7], agglomerative [8, 9], mean shift [10] and affinity propagation [11, 12, 13, 14] while other resources has been identified what kind of parameters are included [15, 16]. We tried to find and demonstrate their behavior when used in a dataset composed of posts written in a low - resources language. For all methods, appropriate data preprocessing is performed, and then the method is used, and the results are analyzed. As a result, a more appropriate data classification and clustering is sought for critical words or terms that frequently appear, as cluster centers. Results and clusters were also manually checked.

The paper starts with an overall introduction, explaining the reasons for selecting these algorithms then, in section II presents a brief literature review. Section III presents the methodology used and steps performed for generating the research results. While section IV presents obtained results with selected algorithms. In the final section of this paper, specifically section V, the conclusions are extracted, and future work is discussed.

II. LITERATURE REVIEW

Collecting the ideas and thoughts from random users, clients of specific services, expressed on social media, in order to use them and improve the company's services, is something very important and should be appreciated by everyone.

After reviewing the relevant literature, we have not found any materials that discuss or implement

experiments similar to ours for low-resource languages. Therefore, we have not pursued this direction of research.

The opinion of each client is very important to the company in order to improve the quality of its services. But sometimes it's hard to use those opinions because they're usually text-based and unstructured. These text responses, customer feedback, are important to the company, but difficult to analyze and extract real meaning from. The known methods used require input of numerical data and provide quantitative estimates, which shows another difficulty such as numerical expression of textual data [17].

Therefore, in this way we need to emphasize that machine learning classifiers seem to be effective tools, methods for detecting clusters in a huge dataset. In terms of classification, text classification methods work in one of three data types such as supervised, semi-supervised and unsupervised data. The selected methods, for this research, have a wide range of use for all three data types and allow implementation to solve different problems and objectives.

According to [5], authors have operated the first mathematical solution of using flexible spectral algorithm in single and multi-dimensional issues using general linear Hermite processes. Also, they propose techniques to use three different adaptive spectral techniques such as p-adaptive, the moving procedure and scaling one. Also, they have tested by using three keys which are associated with scaling, separation and expansion of spectral order. Using these sub-techniques of spectral in explicitly form makes it possible to control error efficiently especially when it is used spectral algorithm.

There are several research papers which discuss how spectral and other algorithms operate in different issues. Authors in [18], propose a new form of implementation which is based on spectral algorithm and tries to detect communities. This means that, by using nodes inside of a system and by learning continuously exceed all forms proposed so far. Authors explain that nodes learn space features of each of them around in low dimensionality and calculate the similarity to finish the community detection.

Also, agglomerative algorithm is used by [9] and represent a framework implementation which has significant impact especially on the accuracy of traditional agglomerative grouping algorithms. This is divided into two implementation perspectives such as single, complete, group linkage and Ward's which are equal with hierarchical model-based method. While the second perspective shows how variants such as complete-link, Mahalanobis-link, and line-link could be used in an extended form in a case of agglomerative algorithm.

Charles Kumah et al. [19], used printed fabric pattern into full color to make image segmentation by using Mean Shift. The dataset contains 11 plain waives and 1 twill waive cotton fabrics. Each printed pattern consists of 600 x 600 pixels while the results have shown that this algorithm is so good in clustering of segmentation for printed fabric patterns even containing texture and illuminations. Approximately in each execution they have gained 6 to 7 clusters where each of them identifies separately object in that figure while at the end it collects all of segments.

Javier Fumanal-Idocin et al [20] propose a construction for novel affinity functions. Nonetheless, they have measured the performance of several datasets using less than-convex combinations of Affinity functions in different communities. At the same time, a proposed method is designated to collect different social mechanisms in a network which has interaction between used functions. Also, they found very good results in the case of modularity measurements for all datasets and algorithms.

III. METHODOLOGY

In this section, we present the way we applied all the methods, including the implementation parameters as well customization. This helped to execute the selected algorithms in the most appropriate way, according to the input data and data preprocessing. In order to analyze the behavior of the above algorithms for the case of low-resource languages, we used datasets consisting of Facebook posts written in Albanian, and the first dataset refers to the company Vala [1] (a telecommunications company) with 1325 posts, and the second to the company Art Motion[2] (a company that provides TV services), with 550 posts, both from the Republic of Kosovo. In the explanation below are used use X and Y, respectively, for those two companies. To convert the posts from text to vector of numbers, TF-IDF is used. The results below, on the left side correspond to dataset X and on the right to dataset Y.

To increase accuracy, we applied stop words removal and word lemmatization. Lemmatization was applied by manually creating a long list of common words which correspond to the root of specific words such as mbushje, mbushjen, mbushjes, mbushjev and these are turned to mbushje. This is applied for both of datasets. Then, we compiled our set of stop words and included a repository of stop words from research [21]. At the same time, we also used data filtering, removal of outliers (which was created a cluster with minimal comments) as well as principal component analysis (PCA) to reduce some dimensions and focus on more significant dimensions and post's words.

To verify the accuracy of the most suitable number of clusters, we used Silhouette which shows the higher accuracy of certain number of clusters. This has helped us to identify specific values for each algorithm.

In the following part, are shown the experimental results of each of the selected methods, applied on both datasets and for different parameter adjustments.

[1] https://www.facebook.com/valamobile

[2] https://www.facebook.com/artmotion.net

IV. EXPERIMENTAL RESULTS

A. *Spectral algorithm*

Spectral, as one of the unsupervised clustering algorithms and which was used in the study [22], and the analyzing of literature review we have decided to use it and see how much it will be suitable in our case.

By analyzing the method, adjusting the necessary parameters, preparing the input data (data preprocessing) and applying it to both datasets, we have seen that the method gives very good results, and we consider it suitable for languages with low resources. The classification was made into five main clusters (labeled as 0, 1, 2, 3 and 4) applied to dataset X and three main clusters (labeled as 0, 1 and 2) to dataset Y.

In addition, in the case of both datasets, we used specific parameters such as min and max values which means that these are the ranges which adapt the minimum and maximum values used on the datasets, while appropriate value of PCA is tested before to see how they operate in our cases. In our cases, the PCA method has contributed to determine the correct values within the range of values and facilitate the process for algorithm to cluster appropriate clusters.

Therefore, to do it in the right way, we have identified an interval of values that are applied, but in this way, we must be careful not to exceed this interval. So, the interval of values varies approximately 0.00325343 to 0.96325343 for the case of X set of data and 3.17385658 to 8.15436154 for Y set of data. Hence, just a prototype of values for P1 and P2 columns has been presented in table I.

TABLE I. WEIGHT OF COMMENTS SEPARATED IN 2 PRINCIPAL COMPONENTS (P1 AND P2) FOR X AND Y DATASETS.

No of comments	X dataset		No of comments	Y dataset	
N/A	P1	P2	**N/A**	P1	P2
0	0.00325343	0.05183993	0	3.1738565	1.40960012
1	0.00965092	0.07244956	1	4.8683066	3.66278086
2	0.04361599	0.08094468	2	1.1499502	1.47490599
…	…	…	…	…	…
1323	0.02807612	0.06707625	421	3.37407994	4.86895000
1324	0.21093181	0.16381852	422	2.8240604	8.15436154
1325	0.12649087	0.18273131	423	1.8556728	4.11241329

From the obtained results, we recognized the biased results, and in the case of set of data X, most posts were classified in cluster 1, while for set of data Y were in cluster 0. These deviations have increased our dilemmas, but we have explanations for this occurrence. And the explanation is, since how the discussion takes place on social network for example, when someone writes about specific topic, such as price of services, then others comment a lot because it is not suitable for all users and they could agree or disagree, and we needed to check if all comments have been grouped in a specific cluster or algorithm has mis-grouped in any other cluster. For this reason, we manually checked all printed results, while we identified a higher accuracy of clustering each of comment in certain cluster.

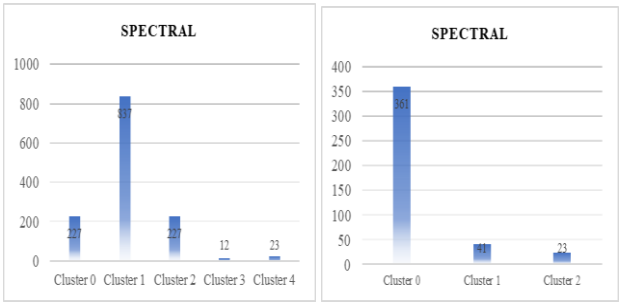


Figure 1. Spectral results using on X and Y datasets.

In Figure 1 we have presented a distribution of comments for both data sets.

To see and prove if the preprocessing is done in the right form, we have used the Silhouette technique to see which cluster has the highest accuracy. By this, we see how many topics are discussed mostly in each dataset.

In the case of dataset X, with the highest accuracy are generated 6 clusters with 69.2%, while for dataset Y are generated 3 clusters with 79.3%.

In table II we have presented in detail for each cluster

TABLE II. CLUSTER ACCURACY OF BOTH DATASETS.

Number of clusters:	Silhouette clusters accuracy of Spectral Clustering	
2	0. 6087469	0.7368314
3	0. 6514980	0.7945238
4	0. 6648371	0.6134574
5	0. 6887702	0.6750648
6	0. 6921797	0.6458574
7	0. 41424305	0.2363203
8	0. 41424305	0.1218392
9	0. 1674287	0.0653652

the accuracy for X and Y datasets while the maximum number of clusters shown is 9. This is because more than 9 clusters have very low percentage of accuracy and we did not present them.

After a whole process of generating clusters, in Figure 2 we have presented visualization of results for certain clusters, while on the left side is for X dataset, and on the right side is for Y dataset.

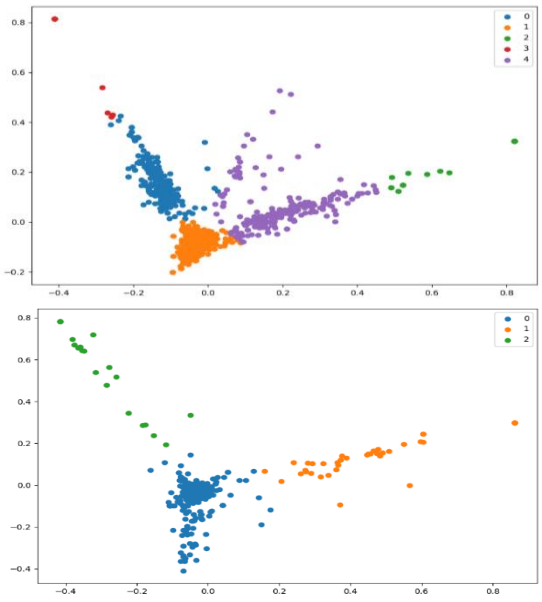


Figure 2. Visualization of data clustering for X and Y datasets.

B. *Agglomerative algorithm*

Agglomeration is hierarchical and works from the bottom up, creating clusters by gathering posts and growing clusters. To analyze its behavior in the case of low-resource languages, we used it to our datasets.

We have applied cosine similarity unlike from other selected algorithms which use PCA. This has contributed directly to increase the performance of visualization at the end stage. While this algorithm has shown a very good ability to link comment threads with others by creating a tree as a whole from the content.

In our case, we measure distance between the comment threads by using Euclidian method because for our dimensionality of data this is the most appropriate. This form of application has helped the algorithm to achieve accuracy and create the most suitable clusters, which means that, from all comments, only 5% have been categorized as outliers (and the conditions for outliers were when they could not link to any cluster while they create cluster by them self's). To be more precise in generating results, we tested different parameters, which correspond to having different results. Some of the parameters we used are in the case of clustering without using lemmatization, Stopwords, vectorization and N-Grams.

In the case when we use all of these parameters, it has generated the best result than in comparison with other tested cases. In Figure 3, we have presented the case, where we use all mentioned parameters for both datasets.

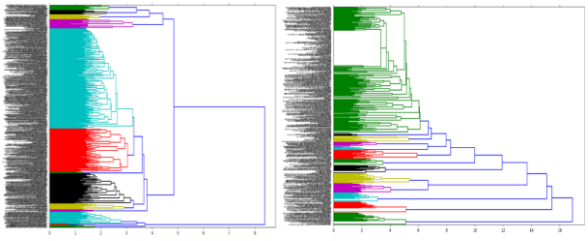


Figure 3. Agglomerative clustering using X and Y sets of data.

Visualizations have differences between two used datasets because grouping is applied into two different contents as it is mentioned in methodology part for whole of the paper. So, in the case of dataset X, a group with teal color has approximately more than 45% of all comments and the other part of the comments thread is on the other groups of the tree linkage. In total, have been generated 6 clusters where each group collect the most appropriate comment in context of sentiment by moving forward to create a tree with the same content as much as it is possible.

In addition, in the case when we use the dataset Y, it has generated 6 groups, but in this case, approximately more than 55% of comment threads are in groups with green color. Grouping of comments in this case Agglomerative algorithm achieved higher accuracy than in the case of X dataset.

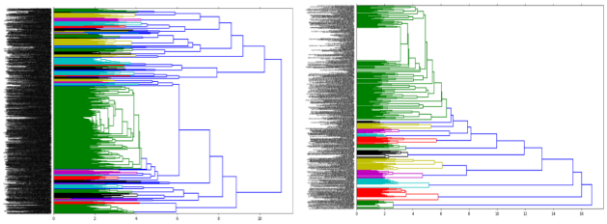


Figure 4. Agglomerative grouping of X and Y sets of data without lemmatization feature.

Results in figure 4, are shown by using other parameters such as Stopwords, N-grams, vectorization, max and min filtering form. So, in this figure, it is not included lemmatization for both cases. Also, here are incorporated N-Grams which it divides terms from 1 to 3, and this has affected by increasing the number of clusters in comparison with other case. This has complicated the process of identifying an appropriate group for each comment thread.

This is because a very large number of groups have been built and determined which one it should be in due to the smallest value from the vectorization. After a manual check of printed results, we identified that in any case within the set of data X and Y, it has clustered in a wrong group any thread of comment while the same comment it shouldn't be there. And this we called as outliers while it hasn't achieved to link in any cluster. So, in this way we have seen that the algorithm has generate a cluster only because they have a term that makes it stand out in terms of writing, while lemma parameter wasn't applied. If a term was there, it has transformed just like "aktivizimi", " aktivizon", and "aktivizu", into a single expression such as "aktivizohet".

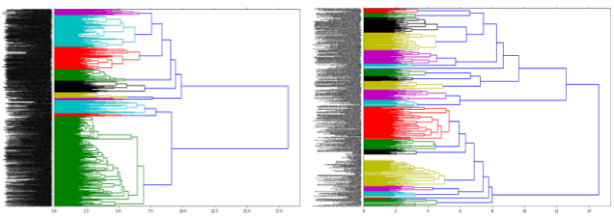


Figure 5. Agglomerative grouping of X and Y sets of data without vectorization, Stopwords and lemmatization.

Figure 5, depicts the final one as commitment in order, wherein we really do not use vectorization, lemma, and the collection of Stopwords. Even though the finding isn't really accurate, the methodology inside the pattern of dendrograms intends to group comment threads far greater than when just lemmas or a list of Stopwords have been used.

In the case of X dataset, most comment threads are classified in a cluster with green color, while in Y dataset, comment threads are grouped more in a cluster with yellow and red color.

C. *Affinity propagation algorithm*

Affinity propagation is based on the implementation of propaganda for each group of content discussed in a dataset. The algorithm works based on the similarity of a comment with others, and this continues until the cluster itself is made. This algorithm is used in different fields and some of them are explained in the research [12]. This algorithm deals with special attributes, analyses and uses datapoints from the dataset to group them by the similarity calculated and found. So, it analyses the comment's content and the relationship between comments using preferences for themselves and other comments.

In the following, we use the same datasets, as in the previous cases, and try to perform parameter tuning on the affinity propagation algorithm and test it in these datasets.

Each comment begins on its own, and to form a cluster, other comments must display the greatest similarity with the given comment. Then, this comment will be joined with the previous one, and so on, until the clusters are formed. This algorithm is based on a preference value on how similar each comment shows itself to others.

Based on the ways that this algorithm operates, the best values for both datasets have been found. This helps to determine for each comment in which cluster it should be, therefore, the values vary from -1, -2, and -3. However, the third case is shown to be the best group by generating better results compared to other cases.

In the case in which the value -2 has been applied, sic clusters generation for X set of data have been affected. Consequently, the distribution of comments is determined based on the number of optimal groups generated that means that the most comments were generated in group 4 (a total of 724), then we have group 5 (213 comments) and so on. Further, the formation of the centroid was calculated based on the highest density of comments at that point, and here the considered terms are “internet, packages and problems, offers/services”.

On the other hand, in the case of dataset Y, the value -1 was applied and 5 groups were generated in total in which 320 comments were grouped in group 2, then in group 0 with 43 comments, and so on. These results for both set of data are shown in figure 6.

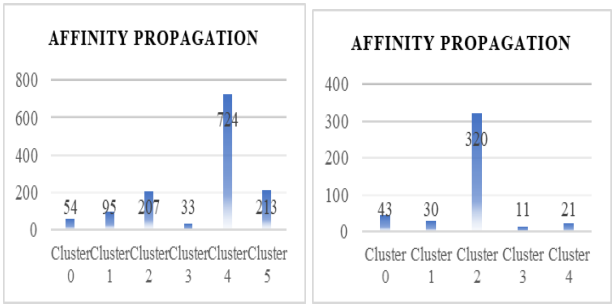


Figure 6. Affinity propagation clustering using X and Y dataset.

The illustration of this is grouping for both cases are shown in figure 7.

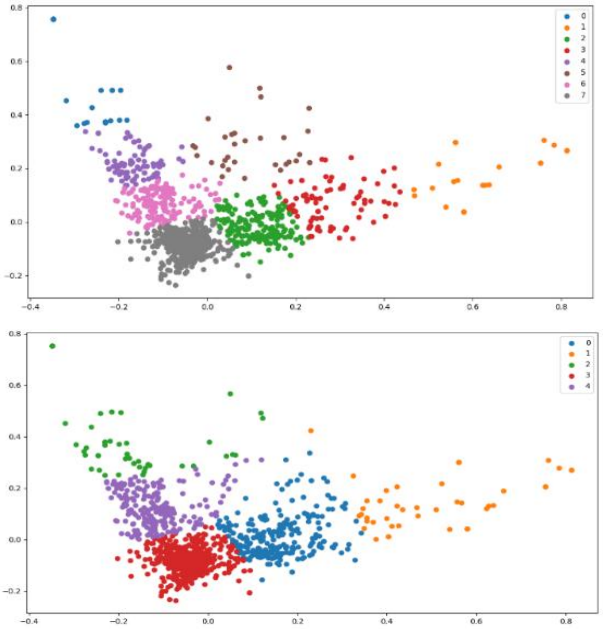
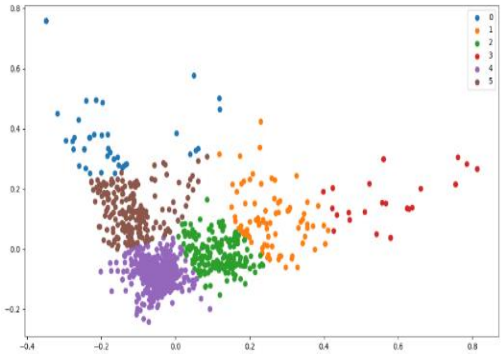


Figure 7. Visualization of results of Affinity propagation clustering using preferences = -1 (in the left side) and -3 (on the right side) of X dataset.

For "preferences = -3" we got 5 clusters, the clusters were created based on keywords such as: problems, Vala, internet, offers and 3G & 4G services, and most of the comments were collected in the same cluster, i.e., 0. Comparing based on the results obtained from all three conducted experiments, we conclude that preferences -2 show better and more accurate results. The “preference = -2” was also shown to be the best choice for the Y dataset as well. The results are shown in Figure 8.



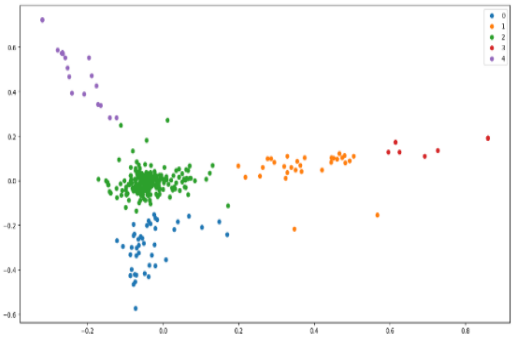


Figure 8. *Visualization of results of Affinity propagation clustering using preferences = -2 for X and Y dataset.*

D. Mean Shift algorithm

The focus of Mean Shift is solely based on the widely held belief that averaging nearly location information results in a greater concentration, and therefore more traditional areas. This algorithm was also applied to find solutions including graphics processing according to [23], [24], in addition to the exhaust system and surface noise removal [25]. A specific used parameter for classifying comment threads by this algorithm is quantile, which is a value that assists to define the data rate to produce the findings. Quantile values have been defined for both datasets and these are applied according to the content of each of them.

At first, we discovered the boundaries, where the classifier works effectively, while it is used to have a specific value that is available to determine the best specific instance of grouping, so this boundary seems to be 0.2 to 0.8 in the scenario of set of data X. In comparison, for set of data Y, the boundary seems to be from 0.321 to 0.621 which produces exactly 8 groups. In this case, when we attempt to increase the value to 0.721, we get 6 groups. Even so, when we increase the value once more, we get only two groups. Considering the above-mentioned range of possibilities (0.02, 0.4, and 0.6 for X set of data), we prefer throughput significance 0.451 as the best optimum value, as well as throughput significance of 0.421 for Y set of data. The number of groups formed by the applied values is presented in figure 9 for both sets of data.

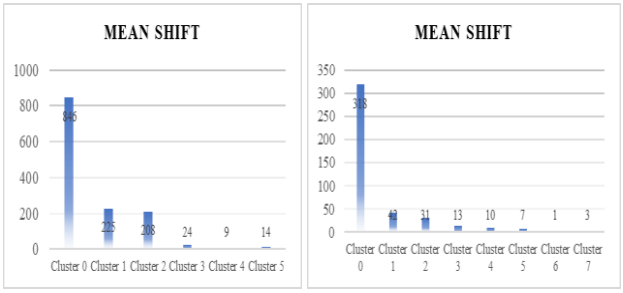


Figure 9. Mean Shift grouping using X and Y datasets.

In figure 10, are shown illustrative forms of how the classification has been generated for all groups for both datasets, applying the parameters mentioned above.

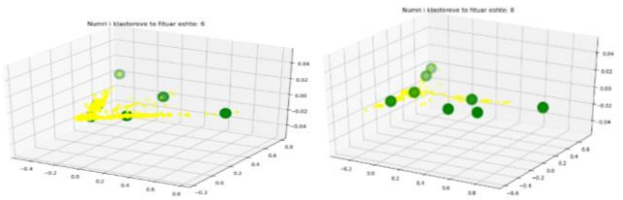


Figure 10. Mean Shift grouping using X and Y datasets.

Then, after several tests we did with values from 0.2 to 0.8, as a constant, it is obtained the following results, with the value of 0.2 are obtained 8 clusters, for 0.3 = 7 clusters, 0.4 = 6 clusters, 0.5 = 7 clusters, 0.6 = 6 clusters, 0.7 = 3 clusters.

After several tests with the increase of the quantile value, in which it exceeds the value of 0.8, then the number of clusters decreased extremely. This proves that the ideal value/limit is between 0.2 and 0.8. Meanwhile, the values with the best results are 0.2, 0.4, and 0.6, all within the above range.

The above quantile values help a lot in identifying the bandwidth of the algorithm (since this algorithm works in that form), and from the value 0.3, it is obtained this value: 0.134282, which subsequently generates 13 clusters. Then we got a 0.4 value from the bandwidth: 0.142802 and it generates 12 clusters, then 0.5 with bandwidth: 0.152266 generates 10 clusters. From this, it seems that as the quantile/bandwidth values increase, the number of clusters decreases.

Further, below we have presented the values obtained during the process of identifying the most suitable cluster:

- 0.182 generates 9 clusters.
- 0.192, 0.211 generate 8 clusters,
- 0.251, 0.281 generate 7 clusters,
- 0.481, 0.451, 0.581 generate 6 clusters,
- 0.681 generates 3 clusters.

Considering the value of 0.251 of bandwidth, the distribution of the clusters was done mostly in cluster 0 with 846 comments. In cluster 1, 224 comments have been collected and in cluster 2 there are 201 comments. Meanwhile, in the other clusters there is a much smaller number of comments, which in cluster 4 and 5 can also be seen as outliers of these results in order not to lose accuracy due to some comments.

V. CONCLUSIONS

The stages of pre-processing and clustering of comments extracted from social networks by analyzing them is now very important for many users, especially companies, as it facilitates and guides them in fulfilling the requests of their users. Therefore, the use of appropriate algorithms to classify each opinion correctly is now essential.

In this paper, we have used two datasets with different contents of low-resource language (in the Albanian language), which, through algorithms, aim to present the groups that are ideal for the topics being discussed. By combining these elements with other algorithms, we aim to determine their perception of the services offered, which is a crucial aspect of our future work. Therefore, to contribute

even a little, through this research, we have selected a total of four algorithms in which, in each of them, the preprocessing stages and the most appropriate parameters are applied, depending on the form of the output that we received from the previous stage.

For example, in the Spectral, Mean Shift, and Affinity Propagation algorithms, the PCA technique is applied, while in Agglomerative cosine similarity. So far, TF-IDF vectorizer, Stopwords list, lemmatization of words, and min and max filtering have been used for all algorithms. Also, for all algorithms, we used Silhouette methods that predict and verify if the results are generated correct from the algorithms used.

The algorithm that generated the best content grouping was shown to be Agglomerative, followed by Affinity Propagation, which once again prove that they are very good in cases where data density is applied. While Spectral generates a very good and suitable number of groups and ideal visualization against the number of groups. This work, in addition to its contribution, also has its limitations, and something like that was access to the most prestigious databases (mainly paid ones) from the literature review part.

Furthermore, based on our thorough research and analysis of existing literature, we have not encountered any prior implementation of similar algorithms in low-resource languages. This posed a significant challenge for us in terms of making comparisons with our own work.

VI. REFERENCES

1 A. Kadriu and L. AbaziA comparison of algorithms for text classification of Albanian news articles,"*Entrenova-Enterprise Research Innovation Conference*Vol. 3 No. 1p. pp. 62–682017

2 A. Kadriu, L. Abazi, and H. AbaziAlbanian text classification: Bag of words model and word analogies*Business Systems Research Journal, doi: 10.2478/bsrj-2019-0006*Vol. 10No. 1p. pp. 74–87Apr. 2019

3 E. Trandafili, N. Kote, and M. Biba,Performance evaluation of text categorization algorithms using an Albanian corpus*Advances in Internet, Data & Web Technologies*p. pp. 537–5472018

4 Stewart, G. Al-Khassaweneh, M. AnAn Implementation of the HDBSCAN* Clustering Algorithm*Appl. Sci., https://doi.org/10.3390/app12052405,* 125pp. 1-212022

5 Khader, M.M.; Adel, M,Modeling and Numerical Simulation for Covering the Fractional COVID-19 Model Using Spectral Collocation-Optimization Algorithm*Fractal and Fractional, https://doi.org/10.3390/fractalfract6070363,* 6363pp. 1-192022

6 Tom Chou, Sihong Shao, Mingtao XiaAdaptive Hermite spectral methods in unbounded domains *Applied Numerical Mathematics* vol. 183p. 201–2202023

7 Kenneth Joseph, Ryan J. Gallagher, Brooke Foucault WellesWho Says What with Whom: Using Bi-Spectral Clustering to Organize and Analyze Social Media Protest Networks*Computational Communication Research*22pp. 153-1742020

8 Arshia Naeem et al.Development of an efficient hierarchical clustering analysis using an agglomerative clustering algorithm*CURRENT SCIENCE*1176pp. 1045-10532019

9 Sepandar D. Kamvar, Dan Klein, Christopher D. ManningInterpreting and Extending Classical Agglomerative Clustering Algorithms using a Model-Based Approach*ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning*pp. 283-2902002

10 S S Pavithra, S Chitrakala, C M BhattSpectral Clustering of Events in Social Media Flood Images based on Multimodal Analysis*7th International Conference on Electrical Energy Systems (ICEES), DOI: 10.1109/ICEES51510.2021.9383727*Chennai, India2021

11 Jia, H., Ding, S., Meng, L. et al.A density-adaptive affinity propagation clustering algorithm based on spectral dimension reduction*Neural Comput & Applic, Springer, https://doi.org/10.1007/s00521-014-1628-7*25p. 1557–15672014

12 Limin WANG, Zhiyuan HAO, Xuming HAN, Ruihong ZHOUGravity Theory-Based Affinity Propagation Clustering Algorithm and Its Applications*Tehnički vjesnik*254pp. 1125-11352018

13 Ge, H. Wang, L. Pan, H. Zhu, Y. Zhao, X. Liu, M.Affinity Propagation Based on Structural Similarity Index and Local Outlier Factor for Hyperspectral Image Clustering*Remote Sensing, MDPI, https://doi.org/10.3390/*141195p. 20222022

14 Limin Wang, Kaiyue Zheng, Xing Tao & Xuming HanAffinity propagation clustering algorithm based on large-scale data-set*InternatIonal Journal of Computers and applICatIons, DOI: 10.1080/1206212X.2018.1425184,* 403pp. 1-72018

15 Campello, R.J.; Moulavi, D.; Sander, J.Density-based Clustering Based on Hierarchical Density Estimates,"*Advances in Knowledge Discovery and Data Mining, Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer*pp. 160-1722013

16 Melvin et al,Visualizing correlated motion with HDBSCAN clustering,"*Protein Science*271pp. 62-752017

17 Singh, U., Saraswat, A., Azad, H.K. et al.Towards improving e-commerce customer review analysis for sentiment detection*Sci Rep 12, https://doi.org/10.1038/s41598-022-26432-3,* p. 022022

18 F. Hu, J. Liu, L. Li et al. "Community Detection in Complex Networks using Node2vec with Spectral Clustering*PHYSA A, Elsevier,* pp. 1-302019

19 Charles Kumah et al.,Unsupervised segmentation of printed fabric patterns based on mean shift algorithm*The Journal of The Textile Institute, Taylor and Francis, DOI: 10.1080/00405000.2020.1867413,* pp. 1-92021

20 Javier Fumanal-Idocin et al.Combinations of Affinity Functions for Different Community Detection Algorithms in Social Networks*Conference: Hawai International Conference on System Sciences,* Hawai: Social and Information Networks2022

21 A. Dine GithubGitHub05072018Available https://github.com/ arditdine /albanian-nlp/blob/master/corpus /stopwords/ albanian. 02062022

22 Mërgim H. HOTI & Jaumin AJDARIUnsupervised Clustering of Comments Written in Albanian Language*International Journal of Advanced Computer Science and Applications (IJACSA*128pp. 287-2922021

23 D. DeMenthonSpatio-temporal segmentation of video by hierarchical mean shift analysis*In Statistical Methods in Video Processing Workshop (SMVP)*Copenhagen, Denmark2002

24 S. Paris and F. DurandA topological approach to hierarchical segmentation using mean shift*in Proc. of the 2007 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR'07)*Minneapolis, MN, USA2007

25 Miguel A. Carreira-PerpinanA review of mean-shift algorithms for clustering*ArXiv, abs/1503.00687*pp. 1-282015