Unsupervised Clustering of Comments Written in Albanian Language

Mërgim H. HOTI, Jaumin AJDARI
Faculty of Computer Science, South East European University (SEEU)
Tetovo, Republic of North Macedonia

***Abstract*—Now-a-days, social media and communications in social media have become very important for services providers and those play a key role in service quality improvement as well as in decision making. The services consumers' discussions usually are written in their local languages and extracting important knowledge sometimes is very hard and problematic. In this field the natural language processing techniques are helpful, but different languages have their specifics and difficulties, and some languages are not prosperous enough in the techniques and methods on NLP, especially the local speaking of the language. In this scientific paper, we have tried to solve such a problem for the Albanian language spoken in Kosovo. Namely, for a dataset of the comments, written in Albanian language in Kosovo (local speaking), collected from the social media, by use of unsupervised clustering techniques, to make clustering regarding the topic of discussion in the comment. In this research, the different techniques of text feature extraction (vectorization and others) and clustering algorithms (K-means, Spectral, Agglomerative, etc.), are used with the idea to find and define more appropriate techniques for the Albanian language. In this paper are shown the results of the conducted experiments as well as discussions about what to use in case of the Albanian language and other languages similar or in group with Albanian (those which have a weak NLP).**

Keywords*—Unsupervised clustering; k-means; spectral; agglomerative; vectorization; Albanian language

I. INTRODUCTION

Data flow nowadays is produced in various industries and fields using technology and this is a challenge in itself which requires management in concrete form. Except for electronic sources such as websites and communication forms by using other mediums. So, social networks are one of the main factors that continue to allow all users to produce different information. At the same time, trends show that different companies tend to use these comments/ sentiment by creating a profile for each user in order to suggest their products according to users' activity. This process, as a whole, requires data management which we otherwise call big data. So, most of the social networks that exist and are used in the world have taken the role of big data producer [1] [2] [3].

This paper mainly examines the use of unsupervised clustering algorithms on social media comments written in Albanian language. Furthermore, this paper demonstrates the text analysis process in reviewing the public opinion of services of Vala Telecommunication Company towards a certain brand and presents hidden knowledge (e.g. services, quality and challenges during operation of their services) that can be used for decision making after the text analysis is performed. Through this paper, we will be focused on implementation of several unsupervised clustering algorithms which they have a wide range field implementation even in different field such as [4], [5], [6], where the main purpose is to identify and create clusters by classifying collected data and distinguishing them as content from extracting and presenting concrete results from processed data through different algorithms. According to Smita Agrawal et al. [2], clustering analysis try to identify the groups of objects such that it forms the groups of similar or related objects groups and in difference forms they are not related to the objects in other groups. Also, Alrence Santiago Halibas et al. [7], shows classification by using similarity by using techniques on English language, the same such in our case we used on Albanian language and as content have classification and clustering of extracted data from social network Twitter. Also, they have used preprocessing techniques preparing data in order to display the visualization of the dataset used.

This paper is organized as follows: Section II is a short literature review regarding the topic, then section III the research methodology used. Implementation of several unsupervised clustering algorithms is shown in Section IV. And, finally, the conclusions and findings in Section V.

II. LITERATURE REVIEW

In this section, we briefly introduce the related technologies involved in our algorithms, including preprocessing phases and visualization of gained results, which are widely used by unsupervised clustering. Our proposed implementation form of clustering of sentiment written on Albanian language is based on the improvement of these three algorithms, which will be described in detail in Section IV.

So, extracting semantic relations has been successfully applied and shown in this part. As found in, Alrence Santiago Halibas et al. [7], use K-Means algorithms where it determines the set of *k* clusters and assigns each example to a specific cluster. This is applied by using sentiment of business analytics on Twitter social network. It has been extracted from the dataset by using preprocessing procedures and visualizing the results of the dataset.

Another researcher, Juan Antonio Lossio-Ventura et al. [8], have used the same social network where they used health-related models and document clustering applications on a Twitter composed of two subsets: HPV and Lynch syndrome

Tweets. Also, it uses Calinski-Harabasz index and Silhouette Coefficient to evaluate the accuracy and performance of implemented algorithms.

Liqiao Zhang et al. [9], propose a methodology of analyzing social media by using consumers' opinions. It uses different social media such as (Facebook, Twitter, Sina Weibo, etc.). Also, it includes three different collective classification algorithms (Local classifier-based method, Logistic regression classifier, Naive Bayes classifier) in the experiment and in the last part of their research shows a visualization method of results, they have reached. According to their conclusions, in the experimental part, the Gibbs sampling method with logistic regression classifier as local classifier performs the best among all the CC (collective classifier) algorithms.

Kai Wang et al. [10], proposes an e-commerce product personalized recommendation system based on learning clustering representation. Also, it uses a methodology of users' for a period time such as income, and occupations, interest. To achieve results for their dataset's [six of them with different of content (shopping, entertainment, sport, film, music and business)], it's used several unsupervised algorithms including Gaussian Mixture Model [11], K-means clustering [12], Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [13], K nearest neighbors (KNN) [14], hierarchical clustering (HC) [15], multi-assignment clustering (MAC). But as a part of their research it noted that the KNN method has its limitations in selecting an adjacent object set. So, they used the neighbor factor and time function and leveraged the dynamic selection model to select the adjacent object set. Also, they combine RNN as well as an attention mechanism to design the e-commerce product results and the performance of proposed algorithms, where it shows better results in six types of dataset.

Kristina P. Sinaga and Miin-Shen Yang [16], propose a new schema with a learning framework for the k-means clustering algorithm in that way where it automatically finds an optimal number of clusters without giving any initialization and parameter selection. Some of the points that have been experimented are feature characteristics, number c of clusters, number n of instances and number f of features in 8 different datasets.

Eric K. Tokuda et al. [17] focuses in applying of agglomerative clustering using unimodal and bimodal datasets where it presents the difference of dendrogram visualization and identifying the clusters in dendrogram, The implementation idea of the proposed approach presents the cluster size s and a number of clusters *k*, the dendrogram is first obtained and then analyzed in a bottom-up approach. Clusters are merged until *k* clusters having at least *s* elements are identified for the first time. Since the last cluster merge might generate a cluster having size much larger than *s*, it is checked if the last merge should be undone.

So, we can clearly see that the chosen algorithms in this paper treat mainly language processing cases, so in our case Albanian language is very specific. Also, these algorithms are shown very successfully in several cases and we are convinced that the results obtained represent the concrete situation in this regard.

III. Research Methodology

In this section, we introduce the state-of-the-art of unsupervised clustering algorithms and sentiment analysis taken from social networks, written on Albanian language. The main focus of the algorithms used is K-Means, Spectral and Agglomerative clustering algorithms.

In our paper we will show three algorithms which we have implemented by using sentiment on social networks written on Albanian language. Problems and challenges of text preparation as in any language, also in Albanian language it has its specifics taking into account the fact of writing not only in one standard or dialect. This further complicates the preparation process or preprocessing stages as it is a very important process in this part.

A. Pre-processing of Dataset

The data that we will use in the acquired dataset are taken from social networks, where to do this we used web scraper, which are taken only the comments/ content of various posts. As a case study we take comments made on the official fan page of VALA Telecommunication[1].

Dataset contains a considerable number of comments which include the various services that the company offers including the latest offers, prices, rewards and services that they offer in the framework of their operation. The dataset is UTF-8-"latin1" encoded, since the Albanian alphabet contains some non-ASCII symbols, like ë, ç, Ç, etc. The next step is to normalize the comments, i.e. to change the comments from upper case to lower case. This step can be skipped to evaluate the influence of the normalization. To split the comments into its words and punctuation marks two different tokenizers are used: The Word2Vect and TF-IDF, which treats a simple emoticon as a single word. The next step in the process is to remove Stopwords. We adopted a list of Albanian Stopwords Ardit Dina [18], where we add more keywords to do more valuable for the implementation on Albanian, also, we comparison with others language such as in Andrej Gajduk and Ljupco Kocarev [19], Henríquez C, Guzmán J [20]. We used Stopwords because we think that it can potentially help in improving performance, and the classification accuracy improved. Now the tokens are converted into *n*-grams. To get the best results, we created a list of words which we have used as stemming of repeated words such as "Interneti", "Internetin", "Internetit", and we have extracted just in one word "Internet", this it helped increasing accuracy of algorithms and optimizing it by easy identifying assigned terms.

B. Feature Extraction

In daily life, implementing machine learning techniques on a large dataset is normal and very important to present the accuracy and originality of results. Every day and more of using social networks and associated types of communication media it produces new amounts of data. This it expresses the

[1] Public company in Republic of Kosovo (https://www.facebook.com/valamobile)

need of managing and implementing new feature extraction by adopting adequate language, whether local language or more spoken language. In fact, techniques of using feature extraction represent a part of the dimensionality reduction process, in which an initial set of the raw data is divided and reduced to more manageable groups. So, this helps the process of managing and seeing if the data are correctly divided according to language perspectives, which it makes it easier. The most important valuable element of using feature extraction on large datasets is that they have a large number of variables. And, the number of variables requires a lot of computing resources to process them. So, feature extraction also helps to get the best feature from those big datasets by selecting and combining variables into features, thus, effectively reducing the amount of data. In our case, after we implement two feature extractions where we have achieved very good results by using of TF-IDF and Word2Vect, where, this it proves that these techniques are suitable and produce original data from core dataset.

To implement TF-IDF we have used two forms of representation data, this we have seen as a process of preprocessing phase:

- Dense vectorization, and;
- Spare vectorization.

As a process, first we present the Dense Vect then the Spare Vect. The explanation of Spare Vect. is that it creates (N x N) matrix, which represents $N_{\text{(horizontal line)}}$ (terms number achieved) **x** $N_{\text{(vertical line)}}$ (number of comments/ rows), then it can be extracted in Dense Vect.

Spare Vectorization form it represents results in this way:

```
00  0800  080010000 10 10 dite ... zoti  çdo  çmim  është  është
kosovës…
0  0.0   0.0    0.0  0.0   0.0  ...   0.0  0.0   0.0    0.0    0.0
1  0.0   0.0  0.0  0.0  0.0  ...    0.0  0.0   0.0    0.0     0.0
```

The list is bigger but in this case we didn't present.

Dense Vectorization form it represents results in this way:

(0, 960)	*0.4576438590753533*
(0, 490)	*0.4758177241051434*
(1, 725)	*0.7071067811865475*
(1, 527)	*0.7071067811865475*
(2, 1096)	*0.32606112714163765*
(2, 877)	*0.30343163250680155*

First column (0, 1, 2 …) shows row/ comments number, second column shows the position of the terms inside of the dataset and the last column shows weight of terms which is identified according to row and column sorting.

Regarding the implementation technique of word2vect we have used two sub-techniques such as:

- Continuous Bag of Words (CBoW), and;
- Skip-grams.

CBoW and Skip-grams have extracted similarities between terms which are used in all datasets.

C. Similarities between Terms in Dataset

CBoW:

[('punso', 0.2354489415884018), ('bane', 0.2187425047159195), ('teknikes', 0.20852388441562653), ('korruptuar', 0.20763424038887024), ('perndryshe', 0.19758988916873932), ('met', 0.19453111290931702), ('besimin', 0.19172176718711853), ('virusi', 0.18643076717853546), ('çdo', 0.18365386128425598), ('master', 0.16957449913024902)]

Skip-Gram:

[('ni', 0.38147056102752686), ('tv', 0.3492451012134552), ('keni', 0.3451632261276245), ('për', 0.34493157267570496), ('sms', 0.34084585309028625), ('vales', 0.3397885262966156), ('150', 0.3364603519439697), ('kerka', 0.3312993347644806), ('euro', 0.3179359436035156), ('ofert', 0.3156954348087311)]

While, in this part we have presented the similarities between two same terms and we present the accuracy of these sub-techniques of Word2Vect.

Table I shows the comparison of these two sub-techniques where it has generated better results for Skip-Grams than CBoW, because of accuracy that Skip-Grams has managed to extract is 0.26 while CBoW 0.05.

TABLE I. SUB-TECHNIQUES OF WORD2VECT.

Sub-Techniques of Word2Vect	
CBoW	**Skip-Grams**
0.058835257	0.2648386

IV. EXPERIMENTAL RESULTS AND ANALYSIS

As we have stated, the focus of this study is to generate and compare the performance of applications using unsupervised clustering algorithms for the sentiment taken from social networks written on Albanian language, where we have specified three of them.

The experiment was developed using the following hardware specifications: Processor: Intel(R) Core(TM) i5, 2.50GHz, RAM 8 GB, System Type: 64-bit Operating System. To see gained results from the extraction data and preprocessing phase, we have used four different criteria of filtering and producing. This, we did to have better results and to see if the data will be presented properly and correctly. This helped the visualization for each implemented algorithm.

The criteria of filtering and production data are:

- With Stopwords & N-Grams;
- With Stopwords & without using N-Grams;
- Without using Stopwords & with N-Grams;
- Without using Stopwords & N-Grams;

In every case, we have achieved a different number of terms, this because when we used N-Grams it divides in more terms such as, "Mbushje", "Mbushja", "Mbushjen" and it make implementation much more complex as a whole.

Achieved results will be presented according to algorithms, which are implemented the techniques explained.

A. K-Means

K-means clustering is a classical clustering method based on data partitioning according to *Hao Yu et al.* [21]. The main idea is to gather the original data into *k* clusters, so that samples with similar attributes are in the same cluster. The main processing procedure is as follows: Firstly, *k* samples are randomly selected from the original data, each sample is taken as the center of *k* clusters, and then the distance between the remaining samples and the *k* center samples is separately calculated, each sample is divided into its nearest center. In total, we achieved three clusters such as Cluster 0, Cluster 1 and Cluster 2. Most of the sentiment is clustered in cluster 0 because it has similar terms between them, than in cluster 2 and the last one is shown in cluster 1. These results are shown in Fig. 1; we have presented a visualization and seen how they stand on this form.

In this way, we have analyzed and shown the mini-framework of K-Means, how it works by using sentiment on Albanian language. The results achieved are satisfactory because of putting centroids in three clusters and every comment which is related to its centroids. The visualization is shown in Fig. 2 where we can clearly see it in three different colors. With yellow representing cluster 1, green is for cluster 2 and the most classified comments from K-Means are clustered in cluster 0 which are in purple.

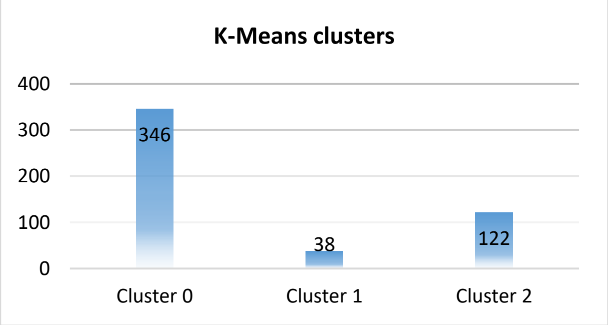

Fig. 1. K-Means Clusters.

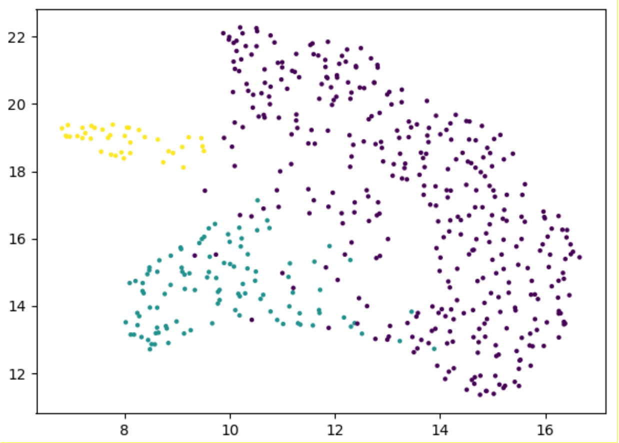

Fig. 2. K-Means Cluster Results.

Our proposed application form of the algorithm is optimized and it is faster than several proposed algorithms in other different researches and examples such as [16], [22] and [23] which are implemented in different cases and datasets. So, the execution time of our proposed form of implementation on Albanian language is 1.662015799999999 seconds. To see and prove, if our visualization is correct and our results correspond with the real statement of implementation we used the silhouette coefficient, to see what kind of results will produce. As we know, silhouette coefficients study and present the separation of distance between the resulting clusters. Also, it computes for each point a width depending on its membership in any cluster. This silhouette width is then an average over all observations according to [8].

This is based on:

$$SC_k = \frac{1}{n} \times \sum_{i}^{n} \frac{b_{i-}a_i}{\max(a_{i,}b_i)}$$

Where, n represents the total number of elements in a cluster, a_i is the average distance between an element i of the cluster and all other elements within the same cluster, b_i represents the average distance between the element i of the cluster and all other elements in the nearest cluster.

Also, we have used silhouette to predict what is the best form of visualization with our dataset. And the results are shown in Fig. 3.

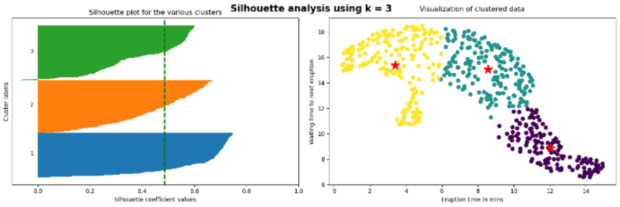

Fig. 3. Silhouette Prediction using Dataset.

While, accuracy is identified with two, three and four clusters presented in percentage in Table II:

TABLE II. SILHOUETTE SCORE WITH DIFFERENT NUMBER OF CLUSTERS

Silhouette clusters accuracy	
Number of clusters: 2	0.7108542990876117
Number of clusters: 3	0.6360185043007658
Number of clusters: 4	0.5455599807949814

B. Spectral Clustering

Spectral Clustering is one of the best known unsupervised algorithms, where, it has performed better than many traditional clustering algorithms in many cases, where we mentioned in related work.

Spectral uses the connectivity approach of clustering, wherein, the parts of nodes (i.e. terms it uses) immediately are next to each other, identified in graphs. The term or connectivity form is then mapped to a low-dimensional space that can be easily segregated to form clusters. Spectral algorithms use data from the eigenvalues of the matrices it created by it i.e. Affinity Matrix, Degree Matrix and Laplacian Matrix derived from the graph or the data we use for our experiment. In our case, we have used the best known techniques for spectral algorithms and the number of clusters' in total are three of them. The most common comments classification is cluster 0 where it has achieved 330 from the total of the dataset, then cluster 1 classified 131 comments and the last one cluster 2 classified 45 comments. These results are shown in Fig. 4, graph model of classification comments for each cluster's.

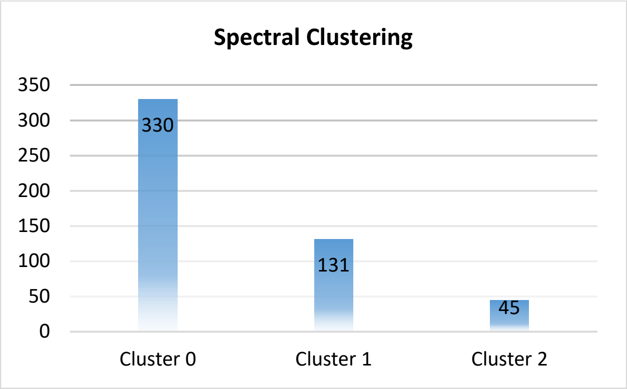

Fig. 4. Spectral Clusters.

Also, spectral clustering visualizes three different colors, which represent each cluster in Fig. 5. For cluster 0 it is purple, green cluster 1 and cluster 2 it's in yellow color. The accuracy of related comment classification with colors are separated very well because the possibility of error is very small, whereas, it seems only three comments which are not classified as it should be but they are missing. Also, we think that as many comments there are in the dataset, the accuracy will be higher. This is because the algorithm can train itself and identify key terms how to separate for each cluster.

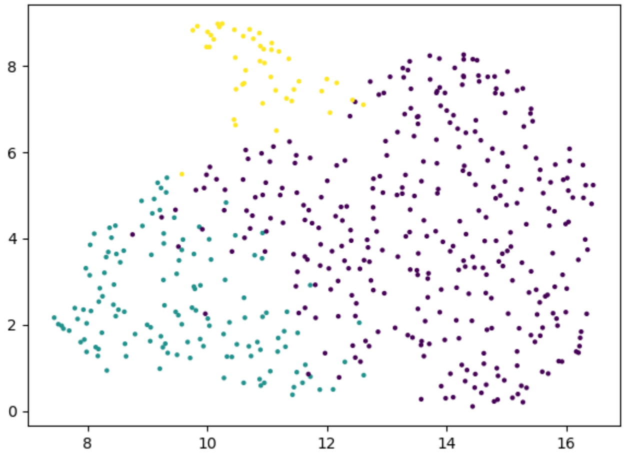

Fig. 5. Spectral Clusters.

C. Agglomerative Clustering

The agglomerative algorithm is part of family algorithms which execute hierarchical clustering. The form of implementation is by grouping objects in clusters based on their terms which are generated by passing preprocessing phases. The main element which increases the accuracy of AC is feature extraction, specifically spare and dense vectorization. The focus of the algorithm is by treating each object as a singleton cluster. In our case, each comment is identified according to the content which is placed in a group of clusters and as seen in Fig. 6, they take on a certain color as a separate cluster. Then, the algorithm continues to compare with other groups of clusters, which, according to the similarities they have managed to make another special group which takes on another special color until the classification of a group in the form of hierarchical clustering as it is presented in Fig. 6. This form of implementation helps algorithms to increase accuracy and present visualization in the best form. So, the results are generated and the algorithm was successful by dividing comments in precisely form for the cluster it should be. The process of clustering is known as hierarchical/ dendrogram model of clustering.

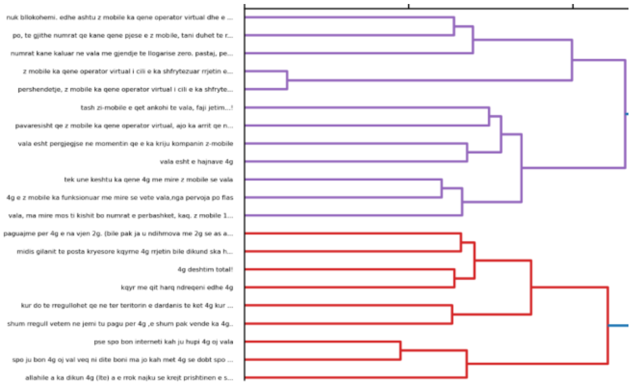
nuk bllokohemi. edhe ashtu z mobile ka qene operator virtual dhe e ...
po, te gjithe numrat qe kane qene pjese e z mobile, tani duhet te r...
numrat kane kaluar ne vala me gjendje te llogarise zero. pastaj, pe...
z mobile ka qene operator virtual i cili e ka shfrytezuar rrjetin e...
pershendetje, z mobile ka qene operator virtual i cili e ka shfryte...
tash zi-mobile e qet ankohi te vala, faji jetim...!
pavaresisht qe z mobile ka qene operator virtual, ajo ka arrit qe n...
vala esht pergjegjse ne momentin qe e ka kriju kompanin z-mobile
vala esht e hajnave 4g
tek une keshtu ka qene 4g me mire z mobile se vala
4g e z mobile ka funksionuar me mire se vete vala,nga pervoja po flas
vala, ma mire mos ti kishit bo numrat e perbashket, kaq. z mobile 1...
paguajme per 4g e na vjen 2g. (bile pak ja u ndihmova me 2g se as a...
midis gilanit te posta kryesore kqyme 4g rrjetin bile dikund ska h...
4g deshtim total!
kqyr me qit harq ndreqeni edhe 4g
kur do te rregullohet qe ne ter teritorin e dardanis te ket 4g kur ...
shum rregull vetem ne jemi tu pagu per 4g ,e shum pak vende ka 4g..
pse spo bon interneti kah ju hupi 4g oj vala
spo ju bon 4g oj val veq ni dite boni ma jo kah met 4g se dobt spo ...
allahile a ka dikun 4g (lte) a e rrok najku se krejt prishtinen e s...

Fig. 6. Agglomerative Clusters.

The list of results it's bigger than it's presented but this is just to understand the idea and results achieved of implementation in our experiment.

V. Conclusion

In this paper, we have applied and tested which of the three selected algorithms is most suitable for the Albanian language. Results obtained are extracted by using the same dataset for three algorithms but the results are different between algorithms. The optimization of each of the algorithms has shown growth and efficiency, as we have presented in this paper the accuracy of the execution of K-Means algorithm dividing into two clusters is 0.71%, three clusters is 0.63% and four clusters is 0.54%. While other algorithms such as Spectral and Agglomerative have shown better results on identification and comparison through terms which create main centroids than groups of their clusters. The best example in our experiments is Agglomerative because the idea of implementation is hierarchical grouping data. So, this is the reason why we have taken only three clusters to see achieved results where this helps to take the average of identifying comments in better form. Finally, we consider that this work is just the first step to improve the accuracy of *k*-means, spectral and agglomerative clustering of dataset/ corpus written in Albanian language. The machine learning models, such as K-Means, Spectral and Agglomerative were used in several different languages but it's the first time on

Albanian according to the best of our knowledge. But, that the techniques and results obtained in this paper help to identify and facilitate the form of use in other content. So, we mention sentiment analysis of consumer in different businesses such as restaurants, hotel, public services, sports and patient impression about the services they receive which they express in several social networks which are written especially on Albanian language. Also, the limitations of this paper are mainly limited academic literature and professional real implementation of surrounding text analytics of social network data. Also the limitations of the work was the lack of research and the form of implementation of the algorithms mentioned for the Albanian language, has shortcomings and needs to be worked on even more in this regard. Future work in this field can also be focused on real-time analytics of social network data streams and improving accuracy and trying to give solutions on distinguishing the dialect of the Albanian language. Also, theoretical analysis and experiments on a benchmark dataset have presented the superiority of our proposed method.

References

[1] J. P. Verma, S. Agrawal, B. Patel, A. Patel, "Big Data Analytics: Challenges and Applications for Text, Audio, video and social media data,," International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI),, vol. 5, no. 1, doi:10.5121/ijscai.2016.5105, pp. 41-51, 2016.

[2] S. Agrawal and A. Patel, "SAG Cluster: An unsupervised graph clustering based on collaborative similarity for community detection in complex networks," Physica A, https://doi.org/10.1016/j.physa.2020.125459, Vols. Journal Pre-proof, p. 6, 2020.

[3] S. Agrawal, A. Patel, "a Study on Graph Storage Database of Nosql,," International Journal on Soft Computing Artificial Intelligence and Applications (IJSCAI),, vol. 5, no. 1, doi:10.5121/ijscai.2016.5104, p. 33–39, 2016.

[4] Shenghan Liu et al., "Unsupervised Clustering-based Non-Coherent Detection for Molecular Communications," IEEE, 10.1109/LCOMM.2020.2985073, vol. Volume: 24, no. Issue: 8, , pp. 1-4, Aug. 2020.

[5] Javier Valdes et al., "Unsupervised grouping of industrial electricity demand profiles: Synthetic profiles for demand-side management applications," Energy, Science direct, Elsevier, https://doi.org/10.1016/j.energy.2020.118962, vol. Volume 215, no. Part A, pp. 1-12, 2021.

[6] Junpeng Tan et al., "Unsupervised Multi-view Clustering by Squeezing Hybrid Knowledge from Cross View and Each View," IEEE TRANSACTIONS ON MULTIMEDIA, pp. 1-14, 2020.

[7] Alrence Santiago Halibas et al., "Application of Text Classification and Clustering of Twitter Data for Business Analytics," in 2018 Majan International Conference (MIC), DOI: 10.1109/MINTC.2018.8363162, Muscat, Oman, 2018.

[8] Juan Antonio Lossio-Ventura et al., "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.

[9] Liqiao Zhang et al., "Predicting and Visualizing Consumer Sentiments in Online Social Media," in IEEE International Conference on e-Business Engineering, Macau, China, 2016.

[10] Kai Wang et al., "E-Commerce Personalized Recommendation Analysis by Deeply-learned Clustering," J. Vis. Commun. Image R., Published by Elsevier Inc., doi: https://doi.org/, Vols. Journal pre-proof, no. Journal pre-proof, pp. 1-7, 2019.

[11] HE, Xiaofei, et al., "Laplacian regularized gaussian mixture model for data clustering," IEEE Transactions on Knowledge and Data Engineering, vol. 23.9, pp. 1406-1418, 2010.

[12] Mingliang Xu, Chunxu Li, Pei Lv*, Lin Nie, Rui Hou,, "An Efficient Method of Crowd Aggregation Computation in Public Areas," IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. (10), pp. 2814-2825, 2018.

[13] Hanxin Chen et al., "Particle Swarm Optimization Algorithm with Mutation Operator for Particle Filter Noise Reduction in Mechanical Fault diagnosis, International," Journal of Pattern Recognition and Artificial Intelligence.

[14] Pasi FRANTI et al., "Fast agglomerative clustering using a k-nearest neighbor graph," IEEE transactions on pattern analysis and machine intelligence, vol. 28, no. 11, pp. 1875-1881, 2006.

[15] Y. Wu et al., "Optimal multimodal fusionfor multimedia data analysis," ACM Multimedia, p. 572–579, 2004.

[16] Kristina P. Sinaga and Miin-Shen Yang, "Unsupervised K-Means Clustering Algorithm," IEEE Access, doi: 10.1109/ACCESS.2020.2988796, vol. Volume 8, pp. 80716 - 80727, 2020.

[17] Tokuda, E.K., Comin, C.H., & Costa, L.D., "Revisiting Agglomerative Clustering," in Computer Vision and Pattern Recognition (cs.CV); Machine Learning (stat.ML), ArXiv, 2020.

[18] A. Dine, "Albanian NLP," GitHUb electronic soruce, (https://github.com/arditdine/albanian-nlp/blob/master/corpus/stopwords/albanian), //, 2018.

[19] Andrej Gajduk and Ljupco Kocarev, "Opinion mining of text documents written in Macedonian language," Computer Science - Computation and Language- arXiv, vol. MASA proceedings, no. 2014arXiv1411.4472G, pp. 1-7, 2014.

[20] Henríquez C, Guzmán J, "A Review of Sentiment Analysis in Spanish," TECCIENCIA, vol. Vol. 12, no. No. 22, pp. p. 35-48, 2017.

[21] Hao Yu et al., "Selfpaced Learning for K-means Clustering Algorithm," Pattern Recognition Letters, Elsevier, vol. Volume 132, pp. p. 69-75, 2020.

[22] Sayar Singh Shekhawat et al., "Twitter sentiment analysis using hybrid Spider Monkey optimization method," Evolutionary Intelligence - Springer-Verlag GmbH Germany, part of Springer Nature, https://doi.org/10.1007/s12065-019-00334-2, vol. Special Edition, no. Special Edition, pp. 1-10, 2020.

[23] Chunhui Yuan and Haitao Yang, "Research on K-Value Selection Method of K-Means Clustering Algorithm," Multidiscplinary Scientific Journal- MDPI, doi:10.3390/j2020016, vol. 2, no. 16, pp. 226-235, 2019.