# EVALUATION OF MACHINE LEARNING TECHNIQUES FOR RESEARCH ARTICLES RECOMMANDATION

*Nuhi Besimi*[1], *Betim Çiço*[2], *Visar Shehu*[1], *Adrian Besimi*[1]

[1] South East European University, [2] Epoka University
e-mails: nuhibesimi@gmail.com, bcico@epoka.edu.al, v.shehu@seeu.edu.mk, a.besimi@seeu.edu.mk
North Macedonia

**Abstract:** Recently, application of machine learning techniques on textual data has become a crucial factor in terms of extracting useful and unknown information from textual documents. This research adds to the machine learning community by evaluating some of the most significant text mining techniques for unsupervised and supervised learning that will supposedly ease the process of literature review for researchers. Furthermore, it evaluates the accuracy and execution time for all the phases of the model by comparing multiple techniques. Results showed that our proposed model can have a positive impact in terms of easing the processing of literature reviews and identify trend topics for a given field. On the other hand, this solution does not perform very well in execution time as the volume of data increases.

**Key words:** recommendation system, supervised learning, unsupervised learning, text mining.

## 1. INTRODUCTION

Recommendation systems are becoming more and more important nowadays in solving and helping human in different tasks. They are continuously used in fields of computer science, finance, medicine, sports and many others, for automating or semi-automating various tasks. They are built on top of different data sets and data structures, starting from various databases (DB), textual data and streaming data. These types of systems can also help researchers to find the most relevant research articles for their research fields. Therefore, in this article we are proposing and evaluating a hybrid model for recommending research articles which is built using both supervised and unsupervised machine learning techniques.

Text mining is one of the most challenging areas in many machine learning applications, mainly because of the nature of data. Textual data is unstructured, and

as such requires additional pre-processing steps. Two of the most important measurements when applying machine learning techniques, and especially when dealing with unstructured textual data are: accuracy and performance. Accuracy issues come as a result of the variety of data, and performance issues as a result of the huge volume. In order to tackle the above-mentioned issues, one must establish a well-defined strategy on storing and processing "Big Data".

Various models and techniques are proposed for overcoming these two challenges, which include distributed environments for parallel processing, cloud platforms, high computing resources, GPU processing, etc. [2, 3, 4]. In one hand, distributed environments for parallel processing are widely used and arguably the most suitable for Big Data [5, 6].

Statistical and analytical algorithms recently have shown very promising results working with structured data. However, analyzing semi-structured and unstructured data is not a straightforward task. Most proposed solutions are ad-hoc solutions which are applied to specific problems.

Researches spent too much time in reading others work and finding research questions. This process requires lots of effort in reading and classifying the relevant papers. The process of literature review is a challenge task for new researchers on different fields of study. Through the proposal in this article we try to ease the process of literature review and speed up the time for defining the research problem.

The goal of this article is to present and evaluate the proposed model for recommending research articles. This represents a novel model which is based on machine learning algorithms and it consists of three phases. In this article we present the overall accuracy of the model and the accuracy of the individual phases. In order to identify and calculate the accuracy, multiple supervised and unsupervised techniques are taken into consideration. These techniques are taken into consideration in order to identify the most suitable ones for our case for each individual phase. The primary hypothesis is that we can ease the process of literature review for researchers and identify trend topics for a given field.

To evaluate existing techniques and propose the best approach, we have tested 15 supervised learning techniques for Phase 1 and 3 unsupervised learning techniques for Phase 2 on two different datasets. For this purpose, we conducted two studies. First, evaluation the accuracy of the unsupervised learning techniques which represents the phase 1 of our model. Second, evaluating the accuracy of the supervised learning techniques which represents the phase 2 of our model.

In Section 2 we are going to present our proposed model, the methods used, the datasets and the environment where all the experiments are performed. Section 3 presents the results where the focus is the accuracy of each step independently and the overall system accuracy. In Section 4 we present the summary and the interpretation of results along with limitations and implications of this system.

## 2. METHODOLOGY

To resolve the problem triggering the research, a suitable plan of actions must be established and then carried on. This section introduces the chosen research strategy, and the specific scientific techniques for data collection. Subsequently, application of the method and research ethics are described. This research, in terms of strategy, follows the experiment approach. ("An experiment is an empirical investigation under controlled conditions designed to examine the properties of, and relationships between, specific factors" [15, p. 74].

### 2.1. Proposed Model

The model that we propose is consisted of three phases. Phase 1 is used to generate clusters and label the clusters. Phase 2 is used to generate a model based on supervised learning techniques. Phase 3 for extracting trend topics for specific fields [13].

Phase 1 represents labelling the clusters (see Fig. 1). For each cluster keywords have been extracted. Based on the keywords a label or list of labels for the clusters are defined. The reason why labelling is needed is to construct a training dataset which is later used to generate a model based on supervised learning algorithms.

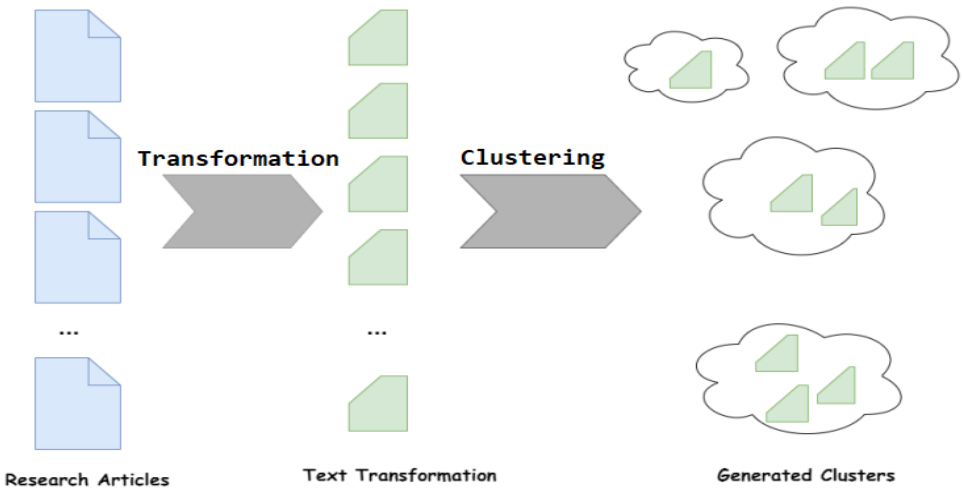


*Figure 1. Phase 1*

Phase 1 is completed now that the training dataset based on the generated clusters have been constructed. For the generated clusters there is a need to also store some metadata, such as: the distance between clusters, the number of clusters, the total number of papers per cluster. The outcomes from Phase 1 are: clusters, distance between clusters, centroids, similarity between papers, top words, outlier clusters.

There are several metrics for assessing the quality of a cluster. The wide range of proposed metric are divided into two categories, external evaluation and internal evaluation. When following the first approach, the clustering is evaluated based on external information such as class labels of the clustered objects. Metrics that belong to this category, compare the results of the algorithms to the actual classes. One simple approach would be to count the number of correctly assigned objects and divide by the total number of clusters. This metrics is known by the name Purity.

When following the second approach, the clustering is evaluated based on the data that was used for clustering. Metrics in this category assess the intra-cluster and inter-cluster distances. An example of such a measure is Dunn Index [7, 8].

In our case, the model relies on the second approach, specifically Dunn Index approach, meaning that the clustering is evaluated based on the data that was used for clustering and no other external information is used.

$$DI = \frac{min\ \delta(Ci,Cj)}{max\ \Delta k} \quad (1)$$

In our proposed model, one of the challenges is the number of clusters. We must identify the right number of k value. Our goal is to detect the optimal number of clusters, in order to ensure high performance on the final phase (recommendation phase). To do so, we rely on the silhouette analysis technique which is a measure of how close each point in one cluster is to the points that belong to another cluster. This measure takes values in the range [-1, 1]. A value of 1 would indicate that an object (point) is far away from the neighboring clusters. Whilst a value of -1 would indicate that the object is very close to the neighboring clusters. Obviously, a desirable silhouette coefficient is the one that is closer to 1. In other words that would mean that the intra-cluster distances are minimized, and the inter-cluster distances are maximized, and we have found the right clusters.

In the case of k-means, k-means++ or k-medoids, we runned the algorithm with different number of values for “k”, compute the silhouette coefficient for each iteration, and identify or report the clustering that produces the highest average Silhouette. The clustering that produces the highest average Silhouette will represent the best value of “k”. In the case of hierarchical clustering, you can choose a view of the dendrogram that provides the highest average Silhouette coefficient.

The outcome from the Phase 1 is n number of clusters. Where n is defined by the Silhouette coefficient. There are two type of clusters which are generated as result of Phase 1, 1) the valid clusters and 2) the outlier clusters. The outlier clusters are the clusters which do not pass the threshold of 25 percentile of the total articles for each cluster. This number is as result of lab experiments with Phase 1 and manual validations of the generated clusters.

The outlier clusters are in special interest for our case. These cluster can represent novel fields of research for a group. In addition to finding these novel

fields, additional metadata needs to be added in order to identify them, like year of publication, references, journal and the similarity with other valid clusters. In contrast to this, there are also outlier clusters which can be ignored, since they contain very few articles in their group (e.x. 1-5 articles) and they do not represent any useful information, except noise from the input dataset.

### 2.2. Phase 2

In the second phase (Fig. 2), a Supervised Learning Technique is used to generate a model based on the training dataset. Crucial about this phase is to choose between various supervised learning algorithms. The training dataset is used as basis for the recommendation system on our research work. The initial results are used to identify the most relevant research articles based on a research interest. The number of recommended research articles depends on the research field and research interest.

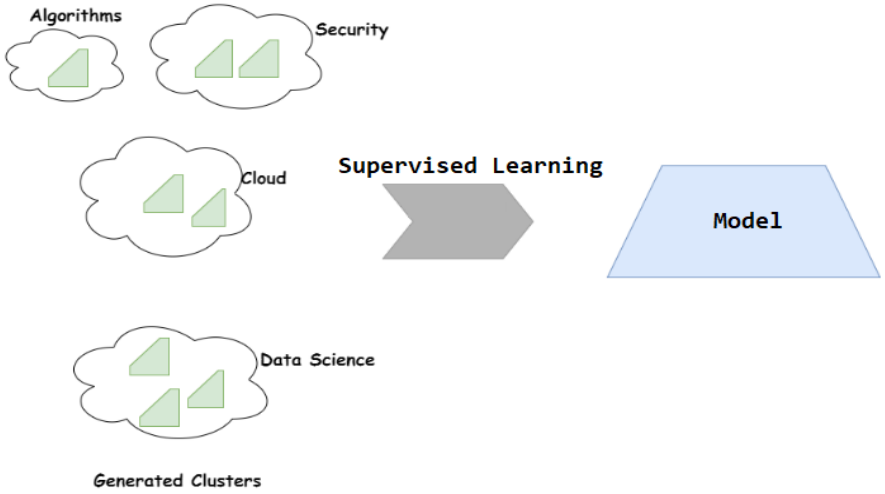


*Figure 2. Phase 2*

The result of Phase 2 is a model based on the training set generated from Phase 1. The Model can be a Decision Tree, Probabilistic Model, Centroids, Neural Network. The aim of our approach is to select the most efficient model based on the literature review and the experiments. This model helps us to: 1) classify new research articles based on their content, 2) recommend research articles based on some search criteria, and 3) query the input dataset for potential research gaps and trend research fields recently.

### 2.3. Datasets

Two datasets were used to test and evaluate the proposed model. First a sample data from a 36GB dataset with research articles in fields of, computer science, neuroscience and biomedical,

Each article was represented by attributes: *id, title, abstract, authors, inCitations, outCitations, year, venue, journalName, journalVolume and sources*. This dataset is used for evaluation the Phase 1 from the model. Second a dataset with 15000 papers in fields of Natural Language Processing, Computer Vision, Methodology, Playing Games, Medical, Graphs and Speech. In Table 1 we present the number of papers for each category.

*Table 1. Dataset organization*

| ***No.*** | ***Category*** | ***# of papers*** |
|---|---|---|
| 1 | Computer Vision | 4872 |
| 2 | Natural Language Processing | 2900 |
| 3 | Medical | 367 |
| 4 | Graphs | 598 |
| 5 | Playing Games | 213 |
| 6 | Speech | 305 |
| 7 | Time Series | 325 |
| 8 | Audio | 88 |
| 9 | Robots | 170 |
| 10 | Computer Codes | 192 |

## 3. RESULTS

### 3.1. Phase 1

First, Dataset 2 was used to measure: (1) the accuracy of the clustering for the K-Means – Table 2 and Fig. 3; (2) the execution time and the impact from the data size – Table 3; (3) efficiency of the silhouette coefficient for the measuring the best split – Table 4; (4) the comparison and the impact of the text representation models on this phase.

*Table 2. Phase 1 Unsupervised Learning Accuracy*

| | ***2 classes*** | ***3 classes*** | ***4 classes*** | ***5 classes*** | ***6 classes*** | ***7 classes*** |
|---|---|---|---|---|---|---|
| TF | 83.4 | 90.3 | 68.2 | 70.1 | 65 | 64.3 |
| TF-IDF | 93.9 | 95.2 | 74.8 | 77.4 | 70.5 | 71.3 |
| TF-IDF and Pre-Processing | 94.2 | 95.5 | 73.2 | 78.2 | 70.5 | 73.2 |
| TF-IDF, Pre-Processing and Title+ (more weight on Title) | 94.2 | 95.5 | 73.2 | 78.2 | 70.5 | 73.2 |

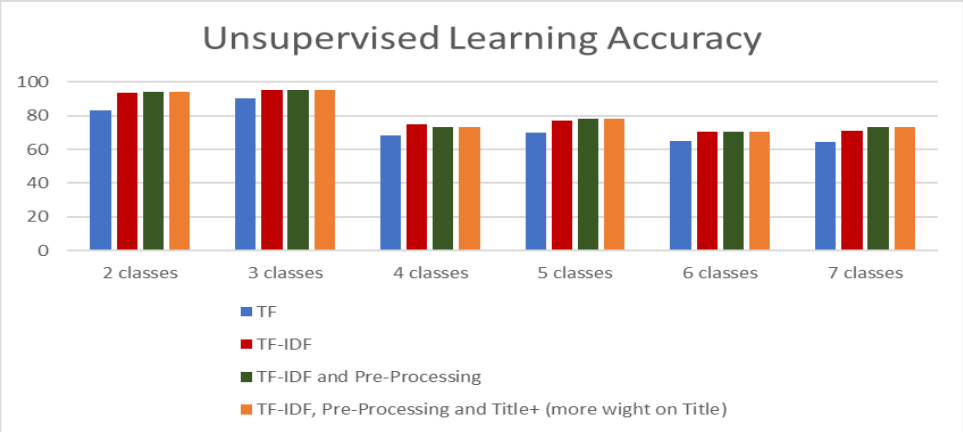


*Figure 3. Phase 1 Clustering Accuracy*

*Table 3. Phase 1 Unsupervised Learning Execution Time in seconds*

| | *100 documents* | *200 documents* | *1000 documents* | *5000 documents* | *10000 documents* | *>20000 documents* |
|---|---|---|---|---|---|---|
| TF | 1.925 | 2.176 | 14.296 | 84.522 | 171 | Memory Error |
| TF-IDF | 1.364 | 2.043 | 22.611 | 92.312 | 195 | Memory Error |
| TF-IDF and Pre-Processing | 3.313 | 3.959 | 27.815 | 115 | 210 | Memory Error |
| TF-IDF, Pre-Processing and Title+ | 3.313 | 3.959 | 30.154 | 120 | 223 | Memory Error |

*Table 4. Efficiency of Silhouette Coefficient (input: 7 clusters)*

| | *2 clusters* | *3 clusters* | *4 clusters* | *5 clusters* | *6 clusters* | *7 clusters* |
|---|---|---|---|---|---|---|
| Silhouette Coefficient | 0.0512 | 0.0626 | 0.0725 | 0.0804 | 0.0893 | 0.0972 |

To give an illustration we performed the same experiments on a sample data from Dataset 1, where only 10000 random selected papers were included. From this experiment in total 37 clusters were generated, where 26 clusters were valid clusters and 11 were outlier clusters – see Table 5.

*Table 5. Generated clusters from Dataset 1*

| **Number of clusters** | **Valid clusters** | **Outlier clusters** |
|---|---|---|
| 37 | 26 | 11 |

The top 6 clusters are presented in the following Table 6.

*Table 6. Top Generated Clusters from Dataset 1*

| **Cluster** | **Papers in cluster** | **Top keywords** |
|---|---|---|
| Cluster 1 | 1257 | Health, disease, medical, evaluation |
| Cluster 2 | 759 | Treatment, brain, therapy, blood |
| Cluster 3 | 364 | Patients, risk, cancer, compared |
| Cluster 4 | 350 | Cell, human, cancer, tumor, DNA |
| Cluster 5 | 312 | System, information, query, strategy, user |
| Cluster 6 | 171 | Algorithm, paper, image, detection |

### 3.2. Phase 2

For the Phase 2 Dataset 2 was used to measure: 1) accuracy of the generated model, 2) the execution time. The following results present an evaluation for phase 2 of our proposed model. For these experiments we have used the dataset 2. The reason why dataset 2 was chosen is because they are already categorized and allow us to estimate the accuracy of different supervised learning techniques. The result showed that different machine learning techniques can lead on different results for textual data. The assessments from Table 7 are graphically interpreted in Fig. 4. Table 8 represents the average accuracy for each technique.

*Table 7. Phase 2 Supervised Learning Accuracy*

| | 2 classes | 3 classes | 4 classes | 5 classes | 6 classes | 7 classes |
|---|---|---|---|---|---|---|
| Random Forest | 0.927 | 0.959 | 0.898 | 0.908 | 0.919 | 0.864 |
| SVM | 0.872 | 0.891 | 0.829 | 0.844 | 0.851 | 0.796 |
| Logistic Regression | 0.854 | 0.891 | 0.813 | 0.853 | 0.860 | 0.796 |
| Decision Tree | 0.872 | 0.931 | 0.835 | 0.866 | 0.864 | 0.823 |
| K-Neighbors (k=5) | 0.890 | 0.897 | 0.781 | 0.655 | 0.300 | 0.233 |
| K-Neighbors (k=15) | 0.872 | 0.877 | 0.851 | 0.825 | 0.800 | 0.759 |
| K-Neighbors (k=30) | 0.872 | 0.884 | 0.835 | 0.857 | 0.860 | 0.793 |
| K-Neighbors (k=50) | 0.836 | 0.918 | 0.840 | 0.844 | 0.860 | 0.812 |
| Naïve Bayes | 0.781 | 0.870 | 0.691 | 0.591 | 0.593 | 0.484 |

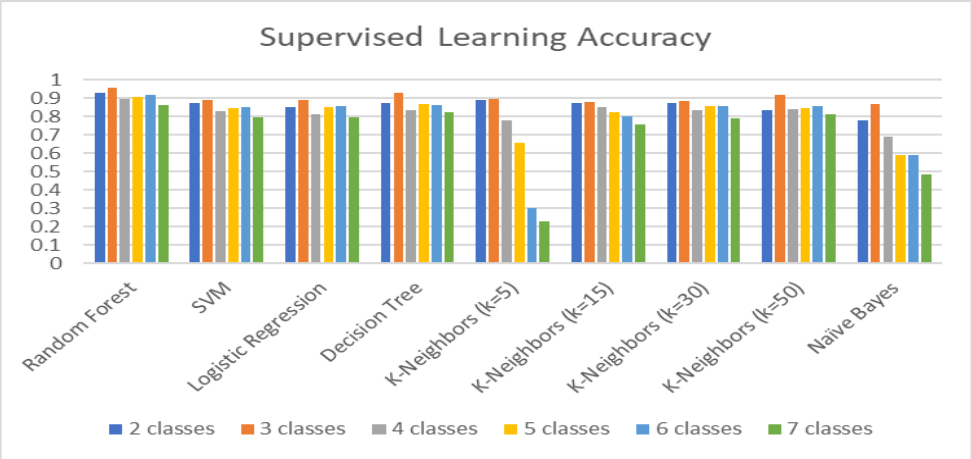


*Figure 4. Phase 2 Supervised Learning Accuracy*

*Table 8. Phase 2 Supervised Learning Average Accuracy*

| # | **Technique** | **Accuracy** |
|---|---|---|
| 1 | Random Forest | 0.86 |
| 2 | SVM | 0.79 |
| 3 | Logistic Regression | 0.79 |
| 4 | K-Neighbors (k=5) | 0.23 |
| 5 | K-Neighbors (k=15) | 0.75 |
| 6 | K-Neighbors (k=30) | 0.79 |
| 7 | K-Neighbors (k=50) | 0.81 |
| 8 | Decision Tree | 0.82 |
| 9 | Neural Network<br>5 layers<br>100 epochs | 0.75 |
| 10 | Naïve Bayes | 0.48 |

## 4. DISCUSSION

In this article we investigated whether a hybrid model based on unsupervised and supervised learning techniques can help researches find relevant articles, identify easier the trend topics and ease the process of the literature review. Our experiments addressed the questions whether 1) the unsupervised learning technique can efficiently generate group of relevant articles by taking into consideration only the title and the abstract of the article, 2) using supervised learning techniques we can generate models which are high accurate on textual data and 3) we can identify trend topics in a given cluster by applying hierarchical clustering and LDA topic modelling. The goal of this research was to evaluate and compare different machine learning techniques for each phase and identify the most accurate ones for our case.

In line with our hypothesis, these results showed that the proposed model can help on easing the process of literature review, but not completely automate it.

Our findings converge with previous findings for Phase 1 (generating the clusters) [9] [10], and with previous findings for the accuracy for Phase 2 (generating the model) [11]. The outcomes in accuracy for Phase 1 and Phase 2 were expected since multiple studies have been performed on different textual datasets where the results were promising. Contrary to the approach used in [12] our research shows that we can recommend research articles also without taking into the consideration the user preferences, activity at all. It can be applied to any dataset without having user or historical data.

Our findings offer a novel perspective to apply/build recommendation system on any dataset of research articles, at different size and at different field of study. You can combine multiple fields in a single dataset and still extract useful information for your research interest. Taken together, our findings on the accuracy and execution time for Phase 1 and Phase 2, and the findings of previous studies point towards that the >85% can be achieved using this approach. Our findings advance our understanding that the application of machine learning can help new researches find relevant articles easier and identify the research gap faster.

Our study has two main limitations. The most important limitation is the size of the initial dataset. If the size of the dataset is big, and you are dealing with a big data problem, then this approach will not be able to completely handle all the phases and provide the expected results. For big datasets, additional enhancements are required on the model to make sure the processing can be done in parallel and distributed fashion. Another limitation is the quality of the input dataset. If the quality of the input dataset is not high this model will not be able to provide helpful end results. Through this approach we are not able to distinguish between a high qualitative and low qualitative input dataset.

## 5. CONCLUSION

In this article we presented and evaluated a model for easing the process of literature review and extracting trend topics from a given dataset of research articles. The results of the three studies showed promising results on accuracy for all the phases of the model. On the other hand, it resulted with increase on execution time and challenges when the dataset size is increasing. Taken together, this offers a novel perspective on recommending research articles, where the user profile or user activity is not taken to consideration.

Future research may extend this work by enhancing the model and removing the limitations on the size of the input dataset. Adding the capabilities for parallel processing on distributed environment. In addition to this, for future we are also

taking into consideration the capability of feeding the model with new research articles and extending/adapting the model and clusters after some period of time.

## REFERENCES

[1] F. K. Putri and J. Kwon, "A Distributed System for Finding High Profit Areas over Big Taxi Trip Data with MognoDB and Spark" in *IEEE International Congress on Big Data (BigData Congress)*, Honolulu, HI, USA, 2017, pp. 533-536.

[2] C.-S. Kim and S.-B. Son, "A Study on Big Data Cluster in Smart Factory using Raspberry-Pi" in *IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA, USA, 2018, pp. 5350-5362.

[3] X. Zhang, "Enabling Effective Utilization of GPUs for Data Management Systems" in *IEEE 33rd International Conference on Data Engineering (ICDE)*, San Diego, CA, USA, 2017, pp. 1612-1615.

[4] Y.-J. Lee, M. Lee, M.-Y. Lee, S. J. Hur and O. Min, "Design of a scalable data stream channel for big data processing" in *17th International Conference on Advanced Communication Technology (ICACT)*, Seoul, South Korea, 2015, pp. 537-540.

[5] Sunil Kumar, Maninder Singh, "A novel clustering technique for efficient clustering of big data in Hadoop Ecosystem" in *Big Data Mining and Analytics*, **4** (vol.2), 2019, pp. 240-247.

[6] A. B. Garay, G. P. Contreras and R. P. Escarcina, "A GH-SOM optimization with SOM labelling and dunn index" in *2011 11th International Conference on Hybrid Intelligent Systems (HIS)*, Melacca, Malaysia, 2011, pp. 572-577.

[7] T. Gupta and S. P. Panda, "Clustering Validation of CLARA and K-Means Using Silhouette & DUNN Measures on Iris Dataset" in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019, pp. 10-13.

[8] F. Liu and L. Xiong, "Survey on text clustering algorithm -Research present situation of text clustering algorithm" in *2011 IEEE 2nd International Conference on Software Engineering and Service Science*, Beijing, China, 2011, pp. 196-199.

[9] X. Wang, "Text clustering based on the improved TFIDF by the iterative algorithm" in *2012 IEEE Symposium on Electrical & Electronics Engineering (EEESYM)*, Kuala Lumpur, Malaysia, 2012, pp. 140-143.

[10] T. B. a. A. v. d. Bosch, "Recommending scientific articles using citeulike" in *Proc. 2008 ACM Conf. Recomm. Syst*, Lausanne, Switzerland, 2008, pp. 287-290.

[11] V. Chaitanya, "Research articles suggestion using topic modelling" in *2017 IEEE 4th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, Port Louis, Mauritius, 2017, pp. 178-182.

[12] N. Besimi, B. Çiço, A. Besimi, "Overview of data mining classification techniques: Traditional vs. parallel/distributed programming models" in *6th Mediterr. Conf. Embed. Comput.*, 2017, pp. 1-4.

[13] N. Besimi, B. Çiço, A. Besimi, "Hybrid solution for research articles recommandation" in *7th Mediterranean Conference on Embedded Computing (MECO)*, Budva, Montenegro, 2019, pp. 1-4.

[14] "Papers with Code," 2019. [Online]. Available: https://paperswithcode.com/sota.

[15] M. Denscombe, The good research guide: for small-scale social research projects, McGraw-Hill Education (UK), 2014.

### *Information about the authors:*

**Nuhi Besimi –** PhD student at South East European University, research interest in fields of Data Mining, Text Mining, Big Data.

**Betim Cico –** Professor at Epoka University, research interest in fields of Data Mining, Data Analysis and Visualization.

**Visar Shehu** - Associate Professor, South East European University at Faculty of Contemporary Sciences and Technologies, research interest in fields of B2B, Data Mining, Web Services

**Adrian Besimi** – Associate Professor, South East European University at Faculty of Contemporary Sciences and Technologies, research interest in fields of B2B, Data Mining, Web Services.

**Manuscript received on 7 February 2020**