Analyzing Graduation Project Ideas by using Machine Learning

Hajar A. Alharbi, Hessa I. Alshaya, Meshaiel M. Alsheail(✉), Mukhlisah H. Koujan
Department of Information Technology, College of Computer, Qassim University,
Buraidah, Saudi Arabia
m.alsheail@qu.edu.sa

Abstract—The graduation projects (GP) are important because it reflects the academic profile and achievement of the students. For many years' graduation projects are done by the information technology department students. Most of these projects have great value, and some were published in scientific journals and international conferences. However, these projects are stored in an archive room haphazardly and there is a very small part of it is a set of electronic PDF files stored on hard disk, which wastes time and effort and cannot benefit from it. However, there is no system to classify and store these projects in a good way that can benefit from them. In this paper, we reviewed some of the best machine learning algorithms to classify text "graduation projects", support vector machine (SVM) algorithm, logistic regression (LR) algorithm, random forest (RF) algorithm, which can deal with an extremely small amount of dataset after comparing these algorithms based on accuracy. We choose the SVM algorithm to classify the projects. Besides, we will mention how to deal with a super small dataset and solve this problem.

Keywords—machine learning, text classification, system analyze, graduation project

1   Introduction

In many years the Information technology department of the girl's section at Qassim University doing many graduation projects. However, all of the efforts are lost because it does not keep it on an electronic system that allowed to take benefits from it, keep it safe, and available all the time. The graduation projects end up stored on a hard disk and printed as hard copy documents and stored randomly in the dust-covered archive room shelves as shown in Figure 1. That makes the search and access on it is so difficult because if anyone needs anything of it or looking for something, need to connect with the responsible of graduation projects then search about it manually, like some cases when the lockdown happened, if some students need an old project, she needs to e-mail the responsible for graduation projects first. Then the responsible will see if the project has an electronic document or not. Also, that is so old way and takes a lot of effort and time also is not with the fact, we are a computer college. Also, this way of dealing with graduation projects deprived us of know the future direction of the IT department

and their alignment with the 2030 vision on the offered projects ideas and it is with trending of the technical world or not. Which makes the process of taking benefit from projects difficult. To solve this problem, we need to create an ML system that allows us to analyze and classified graduation projects. The result well show is the idea of a graduation project it is with trending of the technical world or not is it with the vision 2030 of Saudi Arabia.
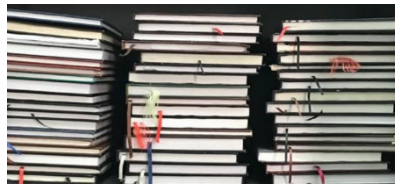
Fig. 1. The random sort for the hard copy of GP

Machine learning (ML) is a system that allows computers to learn without the intervention of a human by using their algorithms [1]. ML is used in many fields that help us in people's lives such as education, business, health care, etc. Also, ML is used in many ways, like recommended systems such as Advertising on YouTube, Netflix, Amazon, etc. Used also on detecting image and voice. It is also used in classification, whether the classification of image, sound, or text. We benefit from ML by using algorithms for classification texts to classify GP. Classification is a very substantial step. By using different supervised algorithms, can classify the text into predefined classes.

Since the number of digital documents is growing so fast, it is become necessary to deal with text classification. Text classification has always been an important topic to research, especially when it needs to work with a large number of texts [2].

Text classification has organized the documents and manages knowledge [3]. From the early period for ML history the usage of text classification technique starts. This technique has often been used in information retrieval systems. Over time with technological development, text classify and documents category start to be used globally in different fields like engineering, healthcare, medicine, psychology, social sciences, law, etc. Also, Text classification is handling for summarize documents [3]. Most of the document categorization and text classification system can be deconstructed into four steps [3] as shown in Figure 2:

- Feature extraction: The documents, in general, are an unorganized dataset, so should be clean it from unnecessary word then applies the methods of Feature extraction.
- Dimension selection: Because the text often has unique words, it has become a problem. To solve this problem, need to use dimension selection.
- Classifier selection: It is to determine the best classification technique, and it is the key step in document classification.
- Evaluations: Understands the performance of the model that used and developed the methods of text classification.
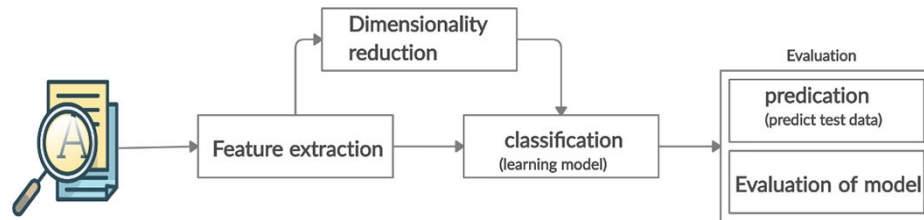
Fig. 2. Steps of categorization

Text classification can be applied into four-level:

1. Document-level: the algorithm well applies the classification for all documents.
2. Paragraph level: the algorithm well applies the classification for a single paragraph in the document.
3. Sentence level: the algorithm well applies the classification for a single sentence.
4. Sub-sentence level: the algorithm well applies the classification for sub-sentence within a sentence [3].

Text classification model: The concept of text classification model and design systems based on ML algorithms. The model will train the classifier accordingly to extant data, after that the classifiers will be tested on the unclassified text. As shown in Figure 3 the following steps are used in the model [4]:

- Text pretreatment.
- Text representation.
- Classifier training.
- Classification.

Text classification is one of the supervised learning tasks. The supervised machine learning is a search and works on undiscovered data expected from a dataset given acknowledged prognostications. It trains on features that are already known (detected) [5]. Some of the famous supervised algorithms used in text classification are:

Support Vector Machine (SVMs): it is a supervised ML algorithm that used kernel concept. The idea of SVMs comes from separating the data by using a line called the hyperplane. It is one of the best algorithms for text classification [5], [6]. The reason for choosing the SVM algorithm as the classifier we use is that after researching and comparing it with other algorithms, we found it is one of the best algorithms for classifying text and the most efficient.
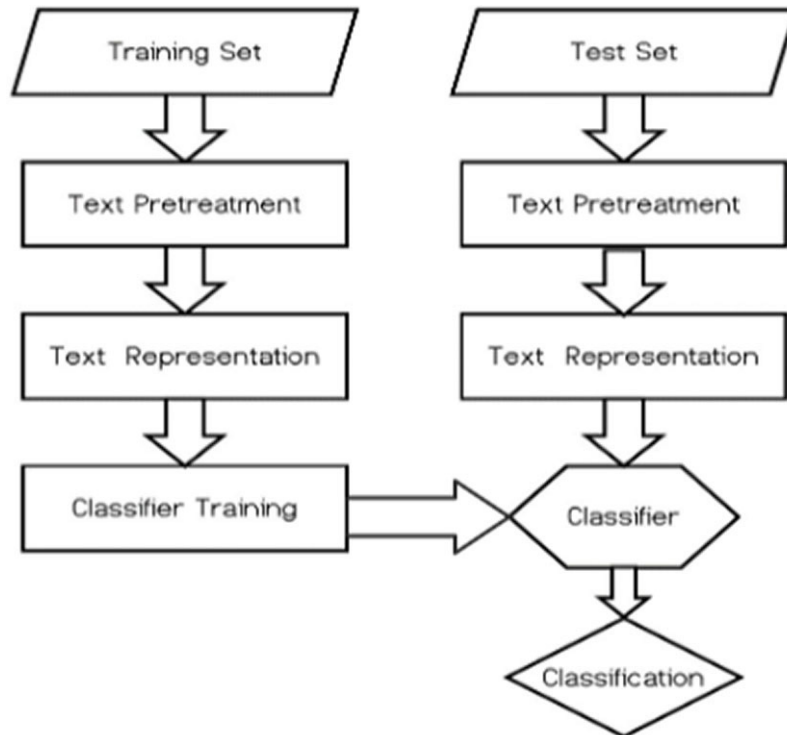
Fig. 3. Text classification model

Random forest (RF): is a flexible and famous machine learning algorithm, it's making the decision tree a categorizer [7]. It does not need prior knowledge, and the categorization accuracy is high without overfitting problems [8]. It's giving great results most of the time, and because of its simplicity and diversity, it's one of the most used algorithms. The RF classifier is made up of a large set of discrete decision trees which work together as a group. Every tree in the random forest produces a class prediction, as well as the class with more votes is becoming the prediction of our model [9].

Logistic Regression (LR) is a robust machine learning algorithm belonging to the Supervised Learning technique. The Logistic Regression algorithm can provide probabilities and classify new data; it uses discrete and continuous datasets. Thus, LR can be used for classifying the explanations by means of diverse kinds of data. It can also determine the most influential variables which are used for the classification [10].

This paper is organized as follows: Section II presents the related work, Section III reports the Research Methodology, Sections IV reports the Analysis and Result Discussions, and finally, Section V concludes the paper.

2   Related work in text classification

Most of the styles are intuitive. However, we invite you to read carefully the brief description below.

So far, there are many related studies and researches about text classification, so in the following we will mention some them.

The authors of [4] design a classification model to classify Chinese news by comparing the precision value, recall, and F-value of three classification algorithms (SVMs, KNN, NB). They found that the SVM algorithm was the highest result while the KNN algorithm and the NB algorithm have the same result. The precision value was 95% for the SVM algorithm, 92% for the NB algorithm, and 92% for the KN algorithm.

The Author in [9] performance of classification methods for text-based data (case in Twitter). ML techniques are used to classify content text training and classification procedures were done several times to reaching the best results. A comparison between Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Logistic Regression (LR), Multinomial Naive Bayes, and Random Forest. Applied these algorithms to classify tweets following a news site to display the matching category. The result showed that SVM outperformed them. Then the two models were chosen, LR and SVM because their results were approximate. Researchers reached the conclusion that SVM is the best algorithms for text classifications.

The authors of [11] used the SVM algorithm to resolve the problem of filing and classification of E-government documents in the E-government information system. The result of classification based on the TP (electronic document classification precision) is 93.7% and on the TN (electronic document misclassification rate) is 6.3%.

In [12] they develop a system to detect any tweet (text) that they consider as cyberbullying and then delete it. They use a number of technologies. First, NLP to identify the Arabic language. Then ML to classify the tweets, whether it was cyberbullying or not. Machine learning models NB and SVM were chosen. The researchers concluded these were the best two methods for classification tasks. A comparison of SVM and NB revealed that SVM performs better NB.

Authors of paper [13] studying the selection feature methods that are generally used with the ML algorithms. Mutual Information (MI) and Chi-square (X2), Term Frequency (TF). They use it with two classifiers Multinomial Naïve Bayes (MNB) and Support Vector Machine (SVM). They tested the methods on a different text dataset. The results showed the average of the classifiers is so close.

This paper [14] used the Naive Bayes algorithm to categorize scientific papers and use LA PDFText tools to extract text from PDF files. The scientific papers are described by defining different sections such as the author, keyword, title, etc. The result is creating an application that makes the user chose a scientific paper then the application will automatically classify it.

This paper [15] used a restricted Boltzmann machine and the kernel-target alignment to develop the restricted Boltzmann machine. The restricted Boltzmann machine will discover the features and the kernel-target alignment will select the features from the set. The method is 10% effective even when the connection between the auxiliary and the target tasks is not apparent.

In this paper [16], they used the Dewey Decimal Classification (DDC) and used a Bag of Word (BOW) method for sorting. To solve the problem of increasing the amount of data and determining massively dispersed data, here is a comparison between four modules to determine the efficiency (Hierarchical Clustering, DDC: Dewey decimal classification, k-Mean Clustering, and SVM: Support Vector Machine), The results pointed out that DDC offer to the most accuracy (75.02%), followed by the Hierarchical models (74.66%), while both K-Mean and SVM offer to the similar accuracy (72.66%). Also, for the time K-Mean Clustering was the best (16.09 seconds).

The need for classification text is increasing because of e-document growth. This paper [17] uses the KM-ELM algorithm to classify the e-document. The proposed system combined two types of machine-learning algorithms. First, a supervised algorithm, which is K-MEANS, and the second, is the unsupervised algorithm, which is extreme learning Machin's (ELM). They use the KM algorithm for feature selection and clustering, and then send it to ELM to use as a training set. After that, the ELM will classify the set. They well using multiple samples and compare each feature for the specified classification for categorization. Moreover, the performance in a different type of dataset. The result is, for the Iris dataset, the accuracy was 85.55%, for Diabetes was 85.7% and for 20 Newsgroups was 86.15%.

In this paper [18], they proposed a new system by combined the SVM algorithm and the KNN algorithm. They apply it on Chinese web pages by using many categories such as finance, health, economics, sports, education, etc. They found the efficiency of the SVM-KNN algorithm together better than when they use the SVM or KNN alone. The efficiency for the automobile category by using the KNN algorithm is 71%, by using the SVM algorithm is 78.8%, and after they use the SVM-KNN algorithm is 79.6%. For the sports category by using the KNN algorithm is 59%, by using the SVM algorithm is 62%, and after they use the SVM-KNN algorithm is 64%.

In this paper [19], they proposed a new method to classify web text. There are studies for text classification techniques by use ca combination for two main methods machine learning (SVM) algorithm and deep learning (CNN). They were combining improved classification accuracy and F-measure. The accuracy of the CNN + SVM algorithm increased from 87.6% to 92.5% and F-measure also increased from 87.9% to 93.2%.

The authors of [20] use the C4.5 to classifying electronic documents. They use it on four types of datasets like 20 Newsgroups, CNAE-9, Reuter-21578, and Twitter. They first classify without a Filtered classifier. The result based on accuracy was 61.04% for 20 Newsgroups, 87.21%, for CNAE-9, 98.13% for Reuter-21578, and 72.6% for Twitter. The result after applying the Filtered classifier on it based on accuracy is 79.81% for 20 Newsgroups, 88.35%, for CNAE-9, 99.35% for Reuter-21578, and 73.4%. Of Twitter which shows an increase in precision. In addition to recent studies as in [21–23].

3 Research methodology

This section describes stages of the proposed methodology in steps: Step 1: Collecting Data, Step 2: Conversion, Step 3: ML algorithm, Step 4: Result. As mentioned before analyzing graduation projects is important to know the popular ideas and projects our college adapts and evaluate whether there is a diversity of ideas or not. In addition, estimate if it fits with the trending technology topics. Figure 4 shows the stages that the project goes through the first stage is collecting the dataset, the second stage is converting the collected dataset from PDF to CSV, the third stage cleaning the dataset, the fourth stage is using the SVM algorithm to classify the graduation projects, the final result we have classified graduation projects.
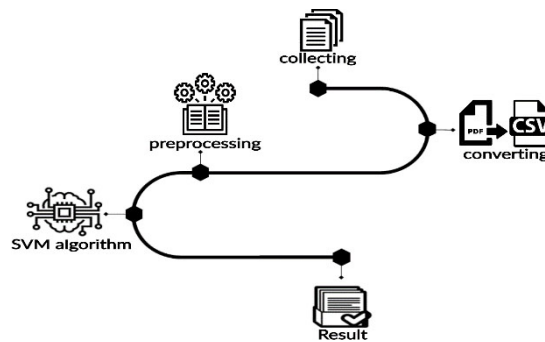

Fig. 4. Implementation stages

For the collecting dataset step, as Figure 5 shows we create our dataset by collecting all the available e-copy, which we got by communicating with the responsible then the implementation begins with collecting e-copy for graduation projects as PDF files. Then converts the PDFs to a CSV file that deals with the ML code. First, we upload a folder that contains the CSV files to train it, Then reading the data files that we uploaded and saving them into new variables, then Put the uploaded files that we saved into variables to loop over it, after that cleaning the dataset by removing the unnecessary data such as numbers, and punctuation and remove it from the dataset, Next splitting the dataset by taking every file in our dataset and split it into two columns the X column have the cleaned data and the Y column the manually classify [24–25].

Merge the dataset we merge all data into one data frame so that we can train our model on it for one time instead of repeating the process over each variable, then splitting the data to training and testing set, Next build the model, then try it on SVM algorithm to classify, lastly, we test our model by uploading new files and apply the prediction function on it so that we can see the results. For the result, the algorithm succeeded in classifying the GP, but the accuracy was not the best because of the very small size of the dataset. We compare it with other algorithms that work with a small amount of dataset, to explore if the reason for low accuracy is from the algorithm or from the dataset.

1.pdf	58,339,057	46,012,290	Microsoft Edge PD...	4/23/2020 7:20 ...	8D82CE8A
2.pdf	16,762,841	14,890,795	Microsoft Edge PD...	4/23/2020 10:5...	0F92C31D
3.pdf	3,509,598	2,375,914	Microsoft Edge PD...	4/23/2020 1:25 ...	C281F26D
4.pdf	9,690,483	8,285,739	Microsoft Edge PD...	4/22/2020 11:1...	1A363EB1
5.pdf	5,165,386	4,610,054	Microsoft Edge PD...	4/20/2020 4:20 ...	CF4D8E2C
6.pdf	8,920,959	8,047,399	Microsoft Edge PD...	4/23/2020 11:2...	75B81764
7.pdf	4,688,299	4,263,738	Microsoft Edge PD...	4/23/2020 3:34 ...	62724F79
8.pdf	1,420,373	1,260,483	Microsoft Edge PD...	4/23/2020 9:11 ...	C532B271
9.pdf	1,275,244	1,045,309	Microsoft Edge PD...	4/24/2020 3:59 ...	10AA75FF
10.pdf	28,570,169	26,615,573	Microsoft Edge PD...	5/12/2020 3:33 ...	F2DF1A76
11.pdf	40,873,167	39,081,179	Microsoft Edge PD...	4/21/2020 8:32 ...	A04F8C65
12.pdf	13,085,205	12,049,719	Microsoft Edge PD...	4/23/2020 8:05 ...	801E077E
13.pdf	12,451,540	11,930,300	Microsoft Edge PD...	4/23/2020 4:43 ...	11AC062A
14.pdf	2,785,672	2,703,741	Microsoft Edge PD...	4/23/2020 8:32 ...	3AE2A698
15.pdf	35,607,235	32,101,327	Microsoft Edge PD...	4/23/2020 8:47 ...	DA83899D
16.pdf	3,120,283	2,773,168	Microsoft Edge PD...	6/12/2020 2:45 ...	0C5401A6
17.pdf	5,873,085	5,174,817	Microsoft Edge PD...	4/23/2020 11:5...	B98210F3
18.pdf	5,070,122	4,719,565	Microsoft Edge PD...	4/23/2020 9:37 ...	13179704
19.pdf	3,833,321	3,055,169	Microsoft Edge PD...	9/28/2020 11:5...	4E64BA7D
20.pdf	2,898,079	2,421,337	Microsoft Edge PD...	9/28/2020 11:5...	7EA4CC4B
21.pdf	2,637,917	2,177,242	Microsoft Edge PD...	9/28/2020 11:5...	34C968E4
22.pdf	7,744,010	7,236,084	Microsoft Edge PD...	9/28/2020 11:5...	D58D67F7
23.pdf	16,725,718	15,334,054	Microsoft Edge PD...	9/28/2020 11:5...	BF011274
24.pdf	2,184,913	1,888,441	Microsoft Edge PD...	9/28/2020 9:41 ...	F59D117A
25.pdf	13,100,194	10,771,705	Microsoft Edge PD...	9/28/2020 9:43 ...	FFD80E27
26.pdf	2,802,321	2,412,725	Microsoft Edge PD...	9/28/2020 9:44 ...	8291BE3C
27.pdf	3,395,043	3,033,108	Microsoft Edge PD...	9/28/2020 9:44	E917D3A1

Fig. 5. Sample of the dataset

4 Analysis and result discussions

4.1 Result

The algorithm succeeded in classifying the GP, but the accuracy was not the best because of the super small size of the dataset, the size of the dataset was (65 rows and 2 Columns) just and that was so difficult to work with. We used a supervised machine learning algorithm for classifying the data. We used the supervised machine learning model called SVM to categorize the GP. The algorithm succeeded in classifying the projects, but with a weak result due to the lack of data to be worked on. The amount of data was supposed to be bigger, but due to the Corona epidemic, we have a severe lack of data. The accuracy rate of the SVM algorithm was 38.3%. Also, we compare it with other algorithms that work with a small amount of dataset as shown in the Table 1 and Figure 6, to explore if the reason for low accuracy is from the algorithm or from the dataset.

Logistic Regression (LR): is a significant machine learning algorithm, which belongs to the Supervised Learning technique. We used this algorithm and the result of the accuracy was 31.64%. Random forest (RF): is a flexible and famous machine learning algorithm, it's making the decision tree a categorizer [7]. It does not need prior knowledge, and the accuracy of the categorization is high without overfitting problems [8]. It's giving great results most of the time, and because of its simplicity and diversity, it's one of the most used algorithms. RF classifier consists of a great number of distinct decision trees that work as a group. Every individual tree in the random forest produces a

classification model, as well as the class with more votes would become the prediction of our model [9]. We used this algorithm and the result of the accuracy was 39.80%.

Table 1. Comparing the algorithms

Algorithms	Accuracy (%)
Support vector machine (SVM)	38.32
Logistic Regression (LR)	31.64
Random forest (RF)	39.80
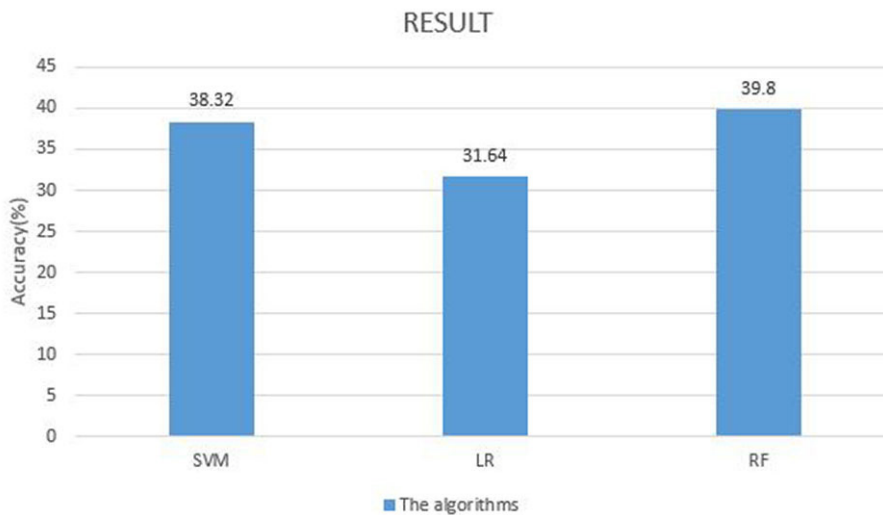
Fig. 6. Comparison accuracy of the algorithms

4.2   Limitation

While working on the project there were some problems, we face it, like:

First: Since most of the graduation projects are only available in the printed format, we were not able to scan it to convert it to e-copy due to the coronavirus pandemic quarantine was imposed and the study system has shifted to online, we did not have access to the archive where the project stored since we did not allow go to campus. Second: The available electronic copies of the graduation projects were in CD format. In addition, while running the CDs we found that some of the CDs worked and others, unfortunately, did not work. The reasons for the damaged CDs might be that some of them are old from many years ago, and some may have been damaged because students borrowed them many times and some of them may have been caused by the wrong way of transferring them from the old university building to the new building. So, the outcome of the electronic versions available from our graduation projects was very few and not enough to build an ideal dataset for the machine learning algorithm. Third: The type of data that we have is specific, so we could not find an available dataset suitable for us.

4.3 Solution we tried

Because we faced a problem of the excessively small amount of the dataset that contained (65 rows and 2 Columns) and that makes the accuracy is very low, so we tried several solutions to solve it: after collecting the dataset and preprocess it we add these steps individually to try to enhance the dataset. First, use a simple classifier because we aim to limit the ability of the model to detect the nonexistent patterns, and reduce the weights like a linear model such as SVM and LR. Second, Detecting the outliers and removing them. The outliers have a substantial effect when it deals with a super small dataset. Third, Use the feature selection. because we have a small and limited dataset this step becomes an absolute step and it will help to deal with this small dataset. Four, Use the bagging classifier with the SVM. We even tried to combine a bagging classifier with an SVM algorithm. Even with all these performance improvements that we tried, it does not affect the accuracy because of the overly small dataset that we have.

5 Conclusions

The aim of this project is to set up a classification system for graduation projects in our department, to provide benefit of technology and help the students who are preparing for the graduation project. The proposed project will help in knowing the department's orientations in its choices for graduation projects, also knowing whether the topics that the college raises for graduation projects are trending with the technical world or not, also help to know about the existing projects with knowledge of their classification and applying to the appropriate field and scientific conferences. Due to our extremely small dataset, we did not get a satisfactory classification result. Among three ML algorithm that we sued even with all these performance improvements that we tried we found that the best performance was SVM with the accuracy rate 38.3%. Since we faced circumstances beyond our control, the COVID-19 pandemic was an obstacle to completing the process of collecting the dataset. In the future, we will convert printed projects into electronic copy and store them on a system that allows us to perform different sorting operations using machine learning algorithms to get the most benefit from this data.

6 References

[1] Boutaba, R., Salahuddin, M. A., Limam, N., Ayoubi, S., Shahriar, N., Estrada-Solano, F., and Caicedo, O. M. (2018). A comprehensive survey on machine learning for networking: evolution, applications and research opportunities. *Journal of Internet Services and Applications*, *9*(1): 1–99. https://doi.org/10.1186/s13174-018-0087-2

[2] Ikonomakis, Emmanouil and Kotsiantis, Sotiris and Tampakas, V. (2005). Text Classification Using Machine Learning Techniques. WSEAS transactions on computers. 4. 966–974.

[3] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., and Brown, D. (2019). Text classification algorithms: A survey. *Information*, *10*(4): 150. https://doi.org/10.3390/info10040150

[4] Miao, F., Zhang, P., Jin, L., and Wu, H. Chinese News Text Classification Based on Machine Learning Algorithm. 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2018, pp. 48–51. https://doi.org/10.1109/IHMSC.2018.10117

[5] Obulesu, O., Mahendra, M., and ThrilokReddy, M. Machine Learning Techniques and Tools: A Survey. In 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), 2018, pp. 605–611. https://doi.org/10.1109/ICIRCA.2018.8597302

[6] Wang, L., Wang, D., and Hao, C. (2017). Intelligent CFAR Detector Based on Support Vector Machine. IEEE Access, 5: 26965–26972. https://doi.org/10.1109/ACCESS.2017.2774262

[7] Xue, D., and Li, F. Research of Text Categorization Model based on Random Forests. In 2015 IEEE International Conference on Computational Intelligence Communication Technology, 2015, pp. 173–176. https://doi.org/10.1109/CICT.2015.101

[8] Guo, Y., Zhou, Y., Hu, X., and Cheng, W. Research on Recommendation of Insurance Products Based on Random Forest. International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), 2019, pp. 308–311. https://doi.org/10.1109/MLBDBI48998.2019.00069

[9] Telnoni, P., Budiawan, R., and Qana'a, M. Comparison of Machine Learning Classification Method on Text-based Case in Twitter. International Conference on ICT for Smart Society (ICISS), 2019, pp. 1–5. https://doi.org/10.1109/ICISS48059.2019.8969850

[10] Logistic regression in machine learning. Accessed: 24.08.2021. [Online]. Available: https://www.javatpoint.com/logistic-regression-in-machine-learning

[11] Mahoto, N. A., Iftikhar, R., Shaikh, A., Asiri, Y., Alghamdi, A., and Rajab, K. (2021). An Intelligent Business Model for Product Price Prediction Using Machine Learning Approach. *Intelligent Automation & Soft Computing*, *29*(3), 147–159. https://doi.org/10.32604/iasc.2021.018944

[12] Pradheep, T., Sheeba, J. I., Yogeshwaran, T., and Pradeep Devaneyan, S. (2017, December). Automatic Multi Model Cyber Bullying Detection from Social Networks. International Conference on Intelligent Computing Systems (ICICS 2017–Dec 15th–16th 2017) organized by Sona College of Technology, Salem, Tamilnadu, India. https://doi.org/10.2139/ssrn.3123710

[13] Chandra, A. Comparison of Feature Selection for Imbalance Text Datasets. International Conference on Information Management and Technology (ICIMTech), 2019, pp. 68–72. https://doi.org/10.1109/ICIMTech.2019.8843773

[14] Rendón-Miranda, J., Arana-Llanes, J., González-Serna, J., and González-Franco, N. Automatic Classification of Scientific Papers in PDF for Populating Ontologies. International Conference on Computational Science and Computational Intelligence, 2014, pp. 319–320. https://doi.org/10.1109/CSCI.2014.153

[15] Zhang, J. Deep Transfer Learning via Restricted Boltzmann Machine for Document Classification. 10th International Conference on Machine Learning and Applications and Workshops, 2011, pp. 323–326. https://doi.org/10.1109/ICMLA.2011.51

[16] Watthananon, J. The relationship of text categorization using Dewey Decimal Classification techniques. 12th International Conference on ICT and Knowledge Engineering, 2014, pp. 72–77. https://doi.org/10.1109/ICTKE.2014.7001538

[17] Neethu, K., Jyothis, T., and Dev, J. Text classification using KM-ELM classifier. International Conference on Circuit, Power and Computing Technologies (ICCPCT), 2016, pp. 1–5. https://doi.org/10.1109/ICCPCT.2016.7530338

[18] Lin, Y., and Wang, J. Research on text classification based on SVM-KNN. 5th International Conference on Software Engineering and Service Science, 2014, IEEE, pp. 842–844. https://doi.org/10.1109/ICSESS.2014.6933697

[19] Wang, Z., and Qu, Z. (2017). Research on Web text classification algorithm based on improved CNN and SVM. 17th International Conference on Communication Technology (ICCT), 2017, IEEE, pp. 1958–1961. https://doi.org/10.1109/ICCT.2017.8359971

[20] Chandrika, G., and Reddy, E. An Efficient Filtered Classifier for Classification of Unseen Test Data in Text Documents. International Conference on Computational Intelligence and Computing Research (ICCIC), 2017, IEEE, pp. 1–4. https://doi.org/10.1109/ICCIC.2017.8524416

[21] Rosba, E., Zubaidah, S., and Mahanal, S. (2021). Digital Mind Map Assisted Group Investigation Learning for College Students' Creativity. International Journal of Interactive Mobile Technologies (iJIM), 15(5): 4–23. https://doi.org/10.3991/ijim.v15i05.18703

[22] Naveed, Q. N., Qureshi, M. R. N., Tairan, N., Mohammad, A., Shaikh, A., Alsayed, A. O. and Alotaibi, F. M. (2020). Evaluating critical success factors in implementing E-learning system using multi-criteria decision-making. *Plos one*, *15*(5), e0231465. https://doi.org/10.1371/journal.pone.0231465

[23] Okuboyejo, S., and Koyejo, O. (2021). Examining Users' Concerns while Using Mobile Learning Apps. *International Journal of Interactive Mobile Technologies (iJIM)*, *15*(15): 47–58. https://doi.org/10.3991/ijim.v15i15.22345

[24] Quasim, M. T., Alhuwaimel, S., Shaikh, A., Asiri, Y., Rajab, K., Farkh, R., and Al Jaloud, K. (2021). An Improved Machine Learning Technique with Effective Heart Disease Prediction System. CMC-COMPUTERS MATERIALS & CONTINUA, *69*(3), 4169–4181. https://doi.org/10.32604/cmc.2021.015984

[25] Platzer, E., and Petrovic, O. (2011). Learning Mobile App Design from User Review Analysis. *International Journal of Interactive Mobile Technologies (iJIM)*, *5*(3): 43–50. https://doi.org/10.3991/ijim.v5i3.1673

7   Authors

Hajer A. Alharbi Graduated with Bachelor's degree in Information Technology. Department of Information Technology, College of Computer, Qassim University, Buraydah, Saudi Arabia. email: 351203615@qu.edu.sa.

Hessa I. Alshaya Graduated with Bachelor's degree in Information Technology. Department of Information Technology, College of Computer, Qassim University, Buraydah, Saudi Arabia. email: 332202751@qu.edu.sa.

Meshaiel M. Alsheail Lecturer at Department of Information Technology, College of Computer, Qassim University, Buraydah, Saudi Arabia Her research interests are Machine learning, User experience, Human computer interaction, E-commerce, and Education. email: m.alsheail@qu.edu.sa.

Mukhlisah H. Koujan Graduated with Bachelor's degree in Information Technology. Department of Information Technology, College of Computer, Qassim University, Buraydah, Saudi Arabia. email: 332220785@qu.edu.sa.