Mesopotamian journal of Big Data
Vol. **(2023)**, 2023, **pp**. 46–52
DOI: https://doi.org/10.58496/MJBD/2023/007 ISSN: 2958-6453
https://mesopotamian.press/journals/index.php/BigData

Research Article

Effectual Text Classification in Data Mining: A Practical Approach

Israa Ezzat Salem[1,*], Alaa Wagih Abdulqader [1] , , Atheel Sabih Shaker[1],

[1]*Computer Techniques Engineering Department, Baghdad College of Economic Sciences University, Baghdad, Iraq*

ARTICLE INFO

Article History
Received 17 Mar 2023
Accepted 2 May 2023
Published 10 May 2023

Keywords
Data mining
Text classification
Techniques
ROC analysis
Data preparation

ABSTRACT

Text classification is the process of setting records into classes that have already been set up based on what they say. It automatically puts texts in natural languages into categories that have already been set up. Text classification is the most crucial part of text retrieval systems, which find texts based on what the user requests, and text understanding systems, which change the text in some way, like by making summaries, answering questions, or pulling out data. Existing algorithms that use supervised learning to classify text automatically need enough examples to learn well. The algorithms for data mining are used to classify texts, as well as a review of the work that has been done on classifying texts. Design/Methodology/Approach: Data mining algorithms that are used to classify texts were talked about, and studies that looked at how these algorithms were used to classify texts were looked at, with a focus on comparative studies. Findings: No classifier can always do the best job because different datasets and situations lead to different classification accuracy. Implications for Real Life: When using data mining algorithms to classify text documents, it's important to keep in mind that the conditions of the data will affect how well the documents are classified. For this reason, the data should be well organized.

1. INTRODUCTION

The data mining approach lets us create scientific discoveries, gain fundamental insights, find new hidden knowledge, discover new patterns, or find patterns of climatic relevance, associations, anomalies, and statistically significant structures or information in data [1-4]. Data mining is employed to classify, cluster, and link the different types of data [5-7]. It uses algorithms for machine learning [8-11]. It has different steps for manipulating data. The first step is the data preparation model, which has different steps for cleaning the data by getting rid of noisy data and filling in missing values, integrating data from different digital resources by getting rid of and reducing redundancy, and then choosing the interesting data to manipulate. Lastly, we set up the data and changed it into other forms. This data preparation model gives us the changed data. Data mining is the name of the second model. This model uses different classification, clustering, and association techniques and algorithms to look for interesting patterns in the data and find them. The user can then see the chosen patterns in different ways. Data mining ends with an evaluation model where users can predict, validate, and interpret the results to confirm or disprove some results or hypotheses [9]. Data mining involves five steps: Data selection, data cleaning, data transformation, pattern evaluation and knowledge presentation and finally decisions / use of discovered knowledge as shown in the Figure 1. Text mining is to handle textual data. Textual data needs to be more structured and clearer, and manipulation is challenging. Text mining is the most suitable method for information exchange. A non-traditional information retrieval strategy is used in text mining to acquire information from a large set of textual documents, which was done by text mining. The figure 2 is elaborated with the process of text mining.

Text classification has been an important application and area of study since the beginning of digital documents [10-12]. Text categorization is becoming more and more important as more and more information is kept in electronic documents [13-16]. User-generated data can contain a lot of useful information, so business and research groups are becoming more interested in how to analyze and get information from it. Text classification is one of the most important parts of natural language processing (NLP) [17-19]. Text classifiers analyze text automatically and put it into a set of tags or categories based on what it says. There are different ways to automatically classify NLP text [20-23]. One way to do this is to use systems that are based on machine learning to automate and speed up the process. There are numerous text documents available in electronic form. More and more are becoming available every day. Such documents represent a massive amount of information that is easily accessible. Seeking value in this huge collection requires organisation; much of the work of organising documents can be automated through data mining. The accuracy and our knowledge of such systems greatly

*Corresponding author. Email: israa.ezzat@baghdadcollege.edu.iq

influence their effectiveness [24]. The task of data mining is to classify records into predefined classes based on their content automatically. Many algorithms have been designed to deal with automatic text classification [25][26]. With the current algorithms, a number of newly established processes are involved in the automation of text classification [27]. The most typical strategies utilised for this objective include linear regression [28], naïve Bayes [29], support vector machine [30], and decision tree [31].

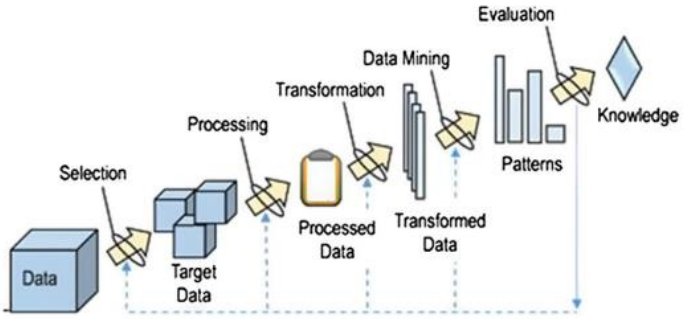

Fig. 1. The main steps in Data mining [32].

NLP involves the development of algorithms and models that enable computers to process human language and the ability to process it. It is an essential tool in data mining, allowing analysts to extract insights and patterns from large amounts of unstructured or random text data. NLP techniques can be operated for many data mining tasks, including text classification, sentiment analysis, and entity recognition. For instance, a machine learning model can be trained to classify text to classify large amounts of unstructured textual data into predefined categories. In view analysis, NLP techniques can be operated to determine the emotional tone of a text, such as whether a review is positive or negative. In addition, entity recognition can help identify and extract information about specific entities, such as individuals or organisations, from large amounts of textual data. NLP techniques are also valuable in medical document mining and radiology reports, where unstructured textual data can be processed into formal computer representations.

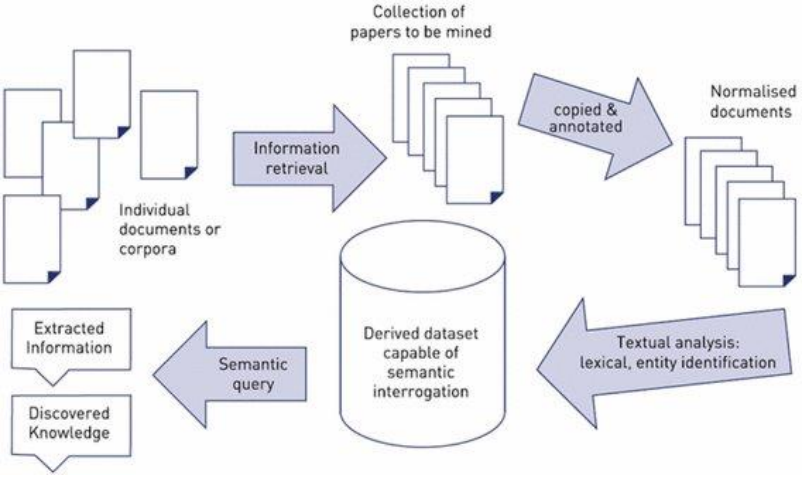

Fig. 2 The process of text mining [33].

Artificial intelligence (AI) has revolutionized various fields, including data mining. Data mining involves analyzing large data sets to extract valuable information and insights [34-36]. The incorporation of AI technologies into data mining has enabled companies and organizations to extract more accurate and valuable insights from their data. AI algorithms can be used to identify patterns and relationships in data that are difficult or impossible for humans to recognize. Machine learning, a subset of artificial intelligence [37][38], enables computers to learn from past experiences and improve their performance over time. As a result, AI algorithms have become essential tools for data mining applications, allowing companies and organizations to identify patterns and trends that would otherwise be difficult to detect. One of the most important benefits of using AI for data mining is its ability to automate the process of identifying patterns and relationships [39][40]. This can save companies and organizations significant time and resources, as they no longer need to rely on human analysts to identify patterns and trends in their data. Alternatively, AI algorithms can be trained to automatically identify patterns and relationships, which can then be employed to create better decisions in a field. Another benefit of using artificial intelligence

in data mining is its ability to handle large amounts of data, especially in the medical area. With the exponential growth of data in recent years, traditional data mining techniques have become increasingly more efficient and desirable. Artificial intelligence algorithms can easily handle large amounts of data of any size, allowing companies and organizations to derive valuable insights from their data. Artificial intelligence techniques are also employed to enhance the accuracy of data mining outcomes. For example, deep learning algorithms can be employed to identify complex patterns in data that are difficult to detect with traditional data mining techniques. As a result, businesses and organizations can make better decisions based on more accurate insights. However, there are also challenges associated with using AI for data mining. One of the biggest challenges is ensuring that the data used to train AI algorithms is accurate and unbiased. If the data used to train AI algorithms is biased, it can lead to inaccurate results and decisions. Additionally, AI algorithms can be complex and challenging to interpret, making it difficult for companies and organizations to understand how they make decisions. The incorporation of AI technologies into data mining has enabled companies and organizations to extract more accurate and valuable insights from their data. By automating the process of identifying patterns and relationships, handling large amounts of data, and improving the accuracy of results, AI has become an essential tool for data mining applications. However, companies and organizations must also be aware of the challenges associated with using AI for data mining, such as ensuring that the data used to train AI algorithms is accurate and unbiased.

Comment on the relationship between information mining and characterization calculation using some open-source information mining tools, such as Rapid Excavator, WEKA, Orange, Knime, and Tangaro. You can see how the four tools work by looking at the accuracy of calculations like Linear regression, Naïve Bayes, Support vector machine, and Decision tree. Testing the data set used in the classification calculation for an Indian liver patient so that the general population can be used as a control [6]. They are talking about the health care field, which has a lot of data and hidden information. By being patient and using this hidden information, good decisions can be made. However, these tests could use less data mining. But there isn't a good tool for analyzing test results and hidden information. So, the system was made by using algorithms for data mining to classify the data and find the heart diseases. Data mining is a solution to many problems in health care. One of these methods is the Linear regression, Naïve Bayes, Support vector machine, and Decision tree algorithm, which is used to diagnose heart diseases. Heart diseases can also be predicted by looking at a few parameters, and a heart disease prediction system (HDPS) is made based on all of the data mining methods.

2. DATA MINING TASKS

Data mining is the process of discovering functional patterns, trends and insights from large, complex datasets. To achieve this goal, various data mining tasks are employed. These data mining tasks can be broadly classified into three categories: descriptive, predictive, and prescriptive. Descriptive data mining tasks are aimed at summarizing or describing the characteristics of a dataset. This type of analysis utilizes techniques such as clustering, association rules, and anomaly detection. On the other hand, predictive data mining tasks involve making predictions or forecasts about future events based on patterns and relationships found in historical data. These tasks commonly use methods such as regression analysis, decision trees, and neural networks to identify trends and forecast future values. Besides, prescriptive data mining tasks employ more sophisticated techniques in order to recommend the best course of action for a given situation. Prescriptive analytics combines the findings from descriptive and predictive analyses to provide actionable recommendations. These techniques include optimization algorithms and simulation modeling. Descriptive data mining tasks can be achieved through various methods such as classification, clustering, regression, summarization, deviation detection and dependency modelling. Classification is a task that involves the categorization of data into predetermined classes or categories based on specific attribute. The tasks associated with data mining can be broken down into several categories depending on how the results of the process are put to use. These categories include [26] [27].

1. Exploratory Data Analysis: This type of analysis consists of nothing more than going over the data without having any specific goals in mind. These methods involve both interaction and visualization.
2. Descriptive Modeling: This modeling technique describes all of the data and also incorporates models that define the general probability distribution of the data, the partitioning of the p-dimensional space into groups, and models that characterize the relationships between the variables.
3. Predictive modeling: This model allows for the value of one variable to be anticipated based on the values of other variables that are already known.

4. Identifying Recurring Norms and Patterns: It is concerned with pattern detection, and its purpose is to identify fraudulent conduct by locating regions of the space defining the various sorts of transactions in which the data points are considerably different from the rest of the space.
5. Retrieval by Content: This method involves locating patterns within the data set that are analogous to the pattern that is of interest to the user. This operation is typically performed on data sets consisting of text and images.

3. EVALUATING TEXT CLASSIFICATION ALGORITHMS

The widget assesses the ability of algorithms (Linear regression, Naïve Bayes, Support vector machine, and Decision tree) to learn. There are different ways to sample, such as using separate test data. The widget can be used in two ways. First, it displays a table with various classifier performance measures, such as classification accuracy and area under the curve. Second, it gives out evaluation effects that other widgets, like ROC Analysis or Confusion Matrix, can be used to determine how well classifiers work. The learner signal is unique in that it can be connected to more than one widget so that the same procedures can be used to test more than one learner. Figure 3 illustrates training models on input texts. Figure 4 presents the results of the models after performing the models training.

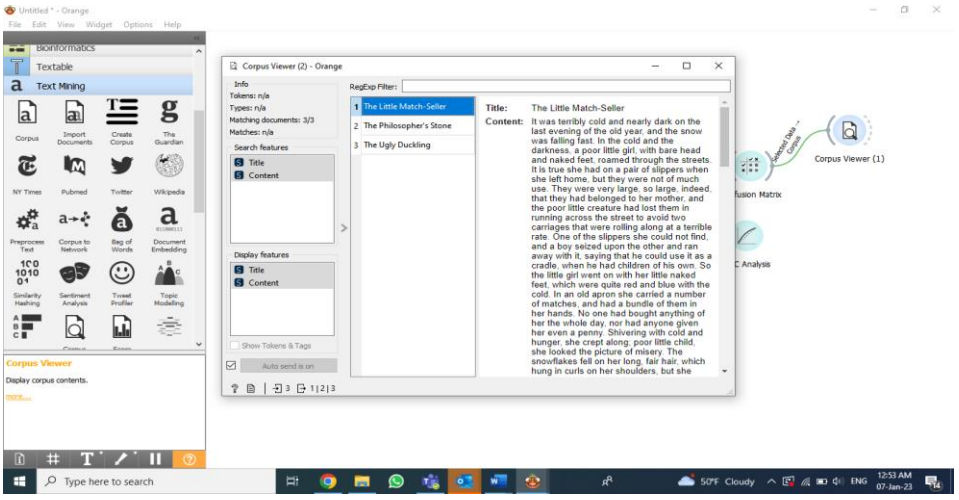

Fig. 3 Model training

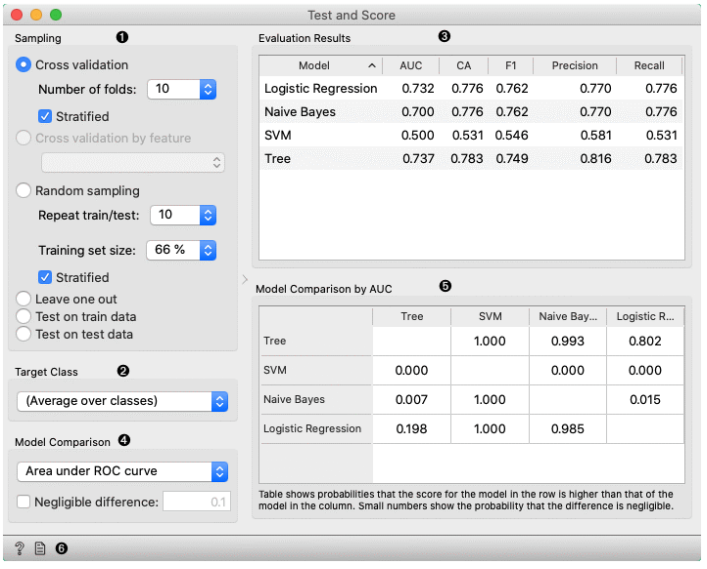

Fig. 4 Test and score.

The confusion matrix compares the actual class to the one that was predicted. Evaluation results: What happened when classification algorithms were put to the test? A subset of data was chosen from the confusion matrix. Data with extra information about whether or not a data instance was chosen. The confusion matrix shows how often the predicted class and the actual class are the same. When an element in the matrix is chosen, the instances that go with it are fed into the output signal. This way, it's easy to see which cases were wrongly categorized and how. Most of the time, Test & Score gives the widget its evaluation results. An example of the schema is shown below (see Figure 5).

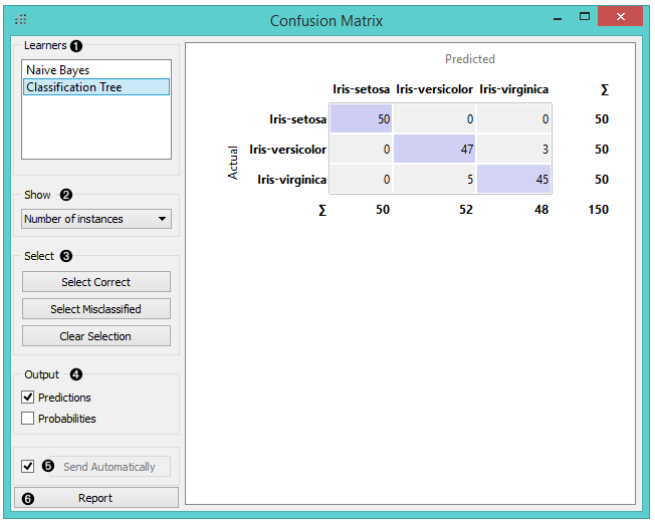

Fig. 5 Confusion matrix.

ROC curve creates a plot that compares a test's true positive rate to its false positive rate. The outcomes of putting classification algorithms through their paces in the evaluation. The ROC curves for the models that were tested, and their matching convex hulls are displayed on the widget. It acts as a standard against which other categorization models can be evaluated. The false-positive rate is defined on the x-axis of the curve as the probability that target = 1 when true value = 0. On the y-axis is the true positive rate, which is defined as the probability that target = 1 when true value = 1. The left-hand border and then the top border of the ROC space should be followed by the curve as closely as possible for the classifier to be as accurate as possible. The widget is also able to select the appropriate classifier and threshold by taking into account the costs associated with false positives and false negatives. Figure 6 presents the results of the ROC analysis, which indicates the ability of the models to separate the data.

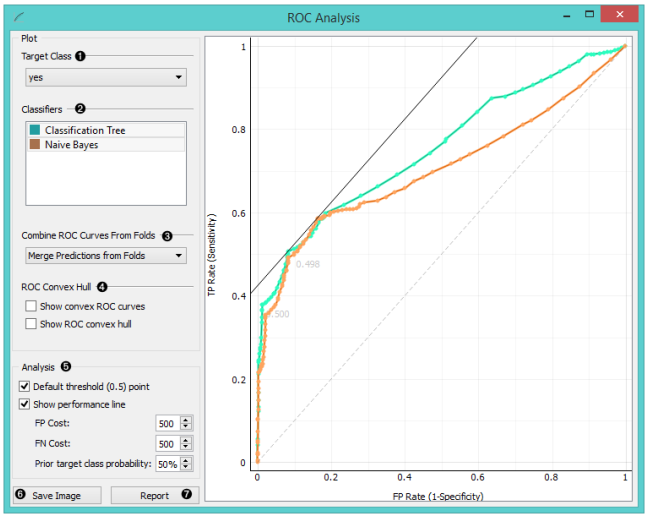

Fig. 6 ROC analysis.

4. CONCLUSION

In this paper, four models (Linear regression, Naïve Bayes, Support vector machine, and Decision tree) are utilised for training and testing the dataset through a set of texts in order to conduct a test between the models and determine which one is better. The tests show that the best model in terms of accuracy scale is the decision tree, which achieved a result of more than 78%, while the worst was the support vector machine, which earned an accuracy of more than 53%.

Conflicts of Interest

The author's paper clearly states that no conflicts of interest exist in relation to the research or its publication.

Funding

The author's paper explicitly states that no funding was received from any institution or sponsor.

Acknowledgment

The author acknowledges the assistance and guidance received from the institution in various aspects of this study.

References

[1] X. Liu, Y. Ding, H. Tang, and F. Xiao, "A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data," Energy and Buildings, vol. 231, pp. 110601, Jan. 2021.

[2] M. Naeem, T. Jamal, J. Diaz-Martinez, S. A. Butt, N. Montesano, et al., "Trends and Future Perspective Challenges in Big Data," in Advances in Intelligent Data Analysis and Applications, Nov. 2021, pp. 309–325.

[3] M. M. Mijwil, K. K. Hiran, R. Doshi, and O. J. Unogwu, "Advancing Construction with IoT and RFID Technology in Civil Engineering: A Technology Review," Al-Salam Journal for Engineering and Technology, vol. 02, no. 02, pp. 54-62, Mar. 2023.

[4] F. Xiao and C. Fan, "Data mining in building automation system for improving building operational performance," Energy and Buildings, vol. 75, pp. 109-118, Jun. 2014.

[5] I. E. Salem, M. M. Mijwil, A. W. Abdulqader, M. M. Ismaeel, A. Alkhazraji, and A. M. Z. Alaabdin, "Introduction to The Data Mining Techniques in Cybersecurity," Mesopotamian journal of cybersecurity, vol. 2022, pp. 28-37, May 2022.

[6] M. I. Al-mashhadani, K. M. Hussein, E. T. Khudir, and M. ilyas, "Sentiment Analysis using Optimized Feature Sets in Different Facebook/Twitter Dataset Domains using Big Data," Iraqi Journal For Computer Science and Mathematics, vol. 3, no. 1, pp. 64–70, Jan. 2022.

[7] O. I. Obaid, "Analysis of H-index and Papers Citation in Computer Science Field using K-Means Clustering Algorithm," Iraqi Journal For Computer Science and Mathematics, vol. 4, no. 2, pp. 1–9, Feb. 2023.

[8] M. M. Mijwil, I. E. Salem, and M. M. Ismaeel, "The Significance of Machine Learning and Deep Learning Techniques in Cybersecurity: A Comprehensive Review," Iraqi Journal For Computer Science and Mathematics, vol. 4, no. 1, pp. 87-101, Jan. 2023.

[9] M. M. Mijwil and I. E. Salem, "Credit Card Fraud Detection in Payment Using Machine Learning Classifiers," Asian Journal of Computer and Information Systems, vol. 8, no. 4, pp. 50-53, Dec. 2020.

[10] A. Dogan and D. Birant, "Machine learning and data mining in manufacturing," Expert Systems with Applications, vol. 166, pp. 114060, Mar. 2021.

[11] A. M. Jiménez-Carvelo, A. González-Casado, M. G. Bagur-González, and L. Cuadros-Rodríguez, "Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity – A review," Food Research International, vol. 122, pp. 25-39, Aug. 2019.

[12] S. Abuzir and Y. Abuzir, "Data Mining For CO2 Emissions Prediction In Italy," Mühendislik Bilimleri ve Araştırmaları Dergisi, vol. 3, no. 1, pp. 59 - 68, 2021.

[13] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, "A Survey on Text Classification Algorithms: From Text to Predictions," Information, vol. 13, no. 2, pp. 83, Feb. 2022.

[14] M. M. Mijwil, K. K. Hiran, R. Doshi, M. Dadhich, A. H. Al-Mistarehi, and I. Bala, "ChatGPT and the Future of Academic Integrity in the Artificial Intelligence Era: A New Frontier," Al-Salam Journal for Engineering and Technology, vol. 2, no. 2, pp. 116-127, Apr. 2023.

[15] C. Audrin and B. Audrin, "Key factors in digital literacy in learning and education: a systematic literature review using text mining," Education and Information Technologies, vol. 27, pp. 7395–7419, Feb. 2022.

[16] [16] A. Jadhav, M. Kaur, and F. Akter, "Evolution of Software Development Effort and Cost Estimation Techniques: Five Decades Study Using Automated Text Mining Approach," Mathematical Problems in Engineering, vol. 2022, no. 5782587, pp. 1-17, May 2022.

[17] M. M. Mijwil, M. Aljanabi, and ChatGPT, "Towards Artificial Intelligence-Based Cybersecurity: The Practices and ChatGPT Generated Ways to Combat Cybercrime," Iraqi Journal For Computer Science and Mathematics, vol. 4, no. 1, pp. 65-70, Jan. 2023.

[18] M. M. Mijwil, M. Aljanabi, and A. H. Ali, "ChatGPT: Exploring the Role of Cybersecurity in the Protection of Medical Information," Mesopotamian journal of cybersecurity, vol. 2023, pp. 18-21, Feb. 2023.

[19] M. Aljanabi and ChatGPT, "ChatGPT: Future Directions and Open possibilities," Mesopotamian Journal of Cybersecurity, vol. 2023, pp. 16–17, Jan. 2023.

[20] V. Dogra, S. Verma, Kavita, P. Chatterjee, J. Shafi, J. Choi, and M. F. Ijaz, "A Complete Process of Text Classification System Using State-of-the-Art NLP Models," Computational Intelligence and Neuroscience, vol. 2022, no. 1883698, pp. 1-26, Jun. 2022.

[21] W. W. Chapman, L. M. Christensen, M. M. Wagner, P. J. Haug, O. Ivanov, J. N. Dowling, and R. T. Olszewski, "Classifying free-text triage chief complaints into syndromic categories with natural language processing," Artificial Intelligence in Medicine, vol. 33, no. 1, pp. 31-40, Jan. 2005.

[22] A. Bhavani and B. S. Kumar, "A Review of State Art of Text Classification Algorithms," Proceedings of International Conference on Computing Methodologies and Communication, Erode, India, 08-10 Apr. 2021, pp. 1-6.

[23] R. Joshi, P. Goel, and R. Joshi, "Deep Learning for Hindi Text Classification: A Comparison," Proceedings of International Conference on Intelligent Human Computer Interaction, Apr. 2020, pp. 94–101.

[24] R. Qasim, W. H. Bangyal, M. A. Alqarni, and A. A. Almazroi, "A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification," Journal of Healthcare Engineering, vol. 2022, no. 3498123, pp. 1-17, Jan. 2022.

[25] J. D. M. Rennie, "Improving Multi-class Text Classification with Naive Bayes," Massachusetts institute of technology — artificial intelligence laboratory, Sep. 2001.

[26] C. Kruengkrai and C. Jaruskulchai, "A parallel learning algorithm for text classification," in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, Jul. 2002, pp. 201–206.

[27] I. E. Ezzat and A. W. Abdulqader, "Predicting Carbon Dioxide Emissions with the Orange Application: An Empirical Analysis," Mesopotamian Journal of Computer Science, vol. 2023, pp. 56–66, Mar. 2023.

[28] Z. He, P. Liu, X. Zhao, X. He, J. Liu, and Y. Mu, "Responses of surface O3 and PM2.5 trends to changes of anthropogenic emissions in summer over Beijing during 2014–2019: A study based on multiple linear regression and WRF-Chem," Science of The Total Environment, vol. 807, no. 2, pp. 150792, Feb. 2022.

[29] E. Donnellan, S. Aslan, G. M. Fastrich, and K. Murayama, "How Are Curiosity and Interest Different? Naïve Bayes Classification of People's Beliefs," Educational Psychology Review, vol. 34, pp. 73–105, Jun. 2021.

[30] M. A. Kadhim and A. M. Radhi, "Heart disease classification using optimized Machine learning algorithms," Iraqi Journal For Computer Science and Mathematics, vol. 4, no. 2, pp. 31–42, Feb. 2023.

[31] M. M. Mijwil, I. E. Salem, and R. A. Abttan, "Utilisation of Machine Learning Techniques in Testing and Training of Different Medical Datasets," Asian Journal of Computer and Information Systems, vol. 9, no. 5, pp. 29-34, Nov. 2021.

[32] J. Z. El Mazouri, M. C. Abounaima, and K. Zenkouar, "Data mining combined to the multicriteria decision analysis for the improvement of road safety: case of France," Journal of Big Data, vol. 6, no. 5, pp. 1-30, Jan. 2019.

[33] P. Zhang, Q. Guo, S. Zhang, and H. H. Wang, "Pattern mining model based on improved neural network and modified genetic algorithm for cloud mobile networks," Cluster Computing, vol. 22, pp. 9651–9660, Nov. 2017.

[34] X. Shu and Y. Yiwan Ye, "Knowledge Discovery: Methods from data mining and machine learning," Social Science Research, vol. 110, pp. 102817, Feb. 2023.

[35] P. M. Seeger, Z. Yahouni, and G. Alpan, "Literature review on using data mining in production planning and scheduling within the context of cyber physical systems," Journal of Industrial Information Integration, vol. 28, pp. 100371, Jul. 2022.

[36] S. V. G. Subrahmanya, D. K. Shetty, V. Patil, B. M. Z. Hameed, R. Paul, et al., "The role of data science in healthcare advancements: applications, benefits, and future prospects," Irish Journal of Medical Science, vol. 191, pp. 1473–1483, Aug. 2021.

[37] M. M. Mijwil, R. Doshi, K. K. Hiran, O. J. Unogwu, and I. Bala, "MobileNetV1-Based Deep Learning Model for Accurate Brain Tumor Classification," Mesopotamian Journal of Computer Science, vol. 2023, pp. 32-41, Mar. 2023.

[38] R. K. Martin, C. Ley, A. Pareek, A. Groll, T. Tischer, and R. Seil, "Artificial intelligence and machine learning: an introduction for orthopaedic surgeons," Knee Surgery, Sports Traumatology, Arthroscopy, vol. 30, pp. 361–364, Sep. 2021.

[39] I. H. Sarker, "AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems," SN Computer Science, vol. 3, no. 158, pp. 1-20, Feb. 2022.

[40] M. Swathy and K. Saruladha, "A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using Machine Learning and Deep Learning techniques," ICT Express, vol. 8, no. 1, pp. 109-116, Mar. 2022.