

Classification of specialities in textual medical reports based on natural language processing and feature selection

Hasanen Abdul-Jawad Hussain Almuhanah, Hawraa Hassan Abbas

Department of Electrical Engineering, College of Engineering, University of Karbala, Karbala, Iraq

Article Info

Article history:

Received Dec 24, 2021

Revised May 17, 2022

Accepted Jun 22, 2022

Keywords:

Electronic medical record
Feature extraction
Feature selection
Machine learning
Natural language processing
Text mining

ABSTRACT

Nowadays, a great deal of detailed information about patients, including disease status, medication history, and side effects, is collected in an electronic format; called an electronic medical record (EMR), and the data serves as a valuable resource for further analysis, diagnosis, and treatment. The huge quantity of detailed patient information in these medical texts produces a huge challenge in terms of processing this data efficiently, however. Machine learning (ML) algorithms, artificial intelligence techniques, and natural language processing tools can have the potential effect of simplifying unstructured data, which could positively affect medical report analysis. Natural language processing (NLP) has recently made huge advances on a variety of tasks. In this paper, an automatic system was thus produced to classify specialist consultant interactions based on patients' medical reports. NLP was used as a pre-processing step on a dataset formed of unstructured medical reports. Feature extraction and selection methods were used to convert the textual reports into sets of features and to extract the most effective features to increase classification accuracy and reduce execution time. Various classification methods were then applied (ML perceptron, logistic regression random forest (RF), and linear support vector classifier (LSVC)). The highest accuracy (99.39%) was achieved in ML-perceptron classification techniques.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Hasanen Abdul-Jawad Hussain Almuhanah

Department of Electrical Engineering, College of Engineering, University of Karbala

Karbala, Iraq

Email: hasanen.a@s.uokerbala.edu.iq

1. INTRODUCTION

Natural language processing (NLP) is a branch of artificial intelligence and machine linguistics that enables computers to access meaning from inputs in the form of natural language or human linguistics. It is used to analyse text or speech, thus allowing machines to access human language [1]. In most well-known languages, such as English, Arabic, Chinese, and Spanish, speech inputs are delivered, whether spoken or written, in human comprehensible forms. The underlying computer language, known as machine language or machine code, is largely incomprehensible to most people, however, as at the lowest levels, machine communication does not utilise words, instead relying on strings of binary data, represented by millions of zeros and ones, which are decoded to produce logical mathematical actions [2].

Medical records include all data collected by medical staff, surgery robots, medical imaging devices, emergency video cameras, and patient monitors in wards, operating theatres, and intensive care units. Healthcare professionals generate many of the important parts of these medical records however, which are recorded in the form of audio files or as text [3].

While the volume and granularity of medical data has grown exponentially over time, most of it remains unused. The main reason for this is that healthcare-related information systems are not equipped to process unstructured data. If the capabilities to interpret, analyse and process unstructured medical data were available, therefore, it is possible that the benefits would be significant both for individual patient treatment and for medical research and public health management purposes [3].

Currently, there is huge interest in applying artificial intelligence in the medical field to improve diagnostics, public health, patient care, and pharmaceutical research, as well as for many other tasks. The success of artificial intelligence (AI) systems in such analysis depends on the availability of datasets and quality of the included data, however, and pre-processing unstructured data is important step towards acquiring efficient information from medical records using AI techniques [3], [4]. Natural language processing, AI, and machine learning (ML) all have the potential to simplify the use of unstructured data; however, it is unlikely that these systems will ever reach a point at which computers can make decisions in critical cases rather than simply supporting the humans who must make those decisions [5], [6].

Data quality: registries must institute data curation to assess and review the quality of data before it is imported into electronic health records (EHRs). This often requires extensive data preparation and cleaning. EHRs are designed to support clinical workflow and to manage the transactions in healthcare as well as the documentation for billing. The conduction of research is not a formal reason for the creation of an EHR data, and EHRs are not designed to systemically collect research-grade longitudinal data. As a result, data captured by EHRs may be of variable quality.

2. RELATED WORKS

Electronic medical records (EMR) are increasingly being used to document and store large amounts of clinical information, yet efficiently and accurately extracting meaningful data from electronic health records is challenging, as a significant portion of clinical information is stored in unstructured, free text. Manual review of this data is thus often necessary which can be time-consuming, error prone, and costly [7]. In terms of using EHRs, researchers in the genomics, clinical imaging, and mobile domains. authors have reviewed the possibility of automating a number of tasks including disease diagnoses, disease prediction, phenotype modelling, disease classification, and developing training representations of medical concepts, such as diseases and medications. Li *et al.* [8] focused on a broad scope of tasks, such as classification and prediction, word embedding, extraction, generation, and similar matters such as question answering, phenotyping, generating knowledge graphs, forming medical dialogue, and supporting multilingual communication and interpretability. They reviewed multiple recent studies that showed how such tasks could be supported by electronic health records and health informatics, concluding that Deep learning methods in the general field of NLP have achieved remarkable success, but that applying them to the field of biomedicine remains challenging due to limited data availability and additional difficulties associated with domain-specific text data.

Rusli *et al.* [9] tried to use text classification in the treatment of people with snakebites to benefit from the patient's benefit when describing the shape of the snake to know the type of snake and thus led to know the necessary treatment for the case. Generating text reports by using a set of questions about the shape of a snake and applying NLP preprocessing, feature extraction, and classification on this text report, and make decision by using four algorithms, support vector machine (SVM), decision trees (DT), naïve Bayes, and k-Nearest Neighbor (k-NN), for training and classification. The results show the classification that gives the highest accuracy was in decision trees algorithm equal to 71.6% with high precision and recall.

One of the promising trends in this field is the development of better knowledge of mining information from unstructured data [10], which is useful when working with a combination of structured and unstructured data to develop better decision making and facilitate broader interpretation. Classification of health-related texts is a special case of text classification. Clinical notes in a particular patient's record may contain a lot of redundancy, due in large part to the documentation habit of many physicians of copying past notes and pasting them into a new note [11].

Hammoud *et al.* [12] presented a new Arabic medical dataset for text classification. The dataset included 2,000 articles over 10 classes (blood, bone, cardiovascular, ear, endocrine, eye, gastrointestinal, immune, liver, and nephrological) of disease. These researchers thus suggested that pre-trained models trained on large, related corpora and fine-tuned to specific datasets yield state-of-the-art results. The bidirectional encoder representations from transformers (BERT)-based model for Arabic biomedical named-entity recognition (ABioNER) model pre-trained on an Arabic medical corpus before fine-tuning on the relevant dataset with the original BERT model produced results of 97.4331 for F1 validation. and 95.9124 in F1 testing, with overall SVMs of 89.1308 and 87.3473, respectively.

Khachidze *et al.* [13] introduced an instrument for classification of medical records based on language, based on 24,855 text records. The documents were classified into three groups (endoscopy, ultrasonography, and X-ray), with 13 subgroups, using two methods; these were SVM and k-NN with feature selection. The results obtained demonstrated that both machine learning methods performed successfully, with SVM operating slightly more effectively based on feature selection. However, at the second stage of classification into subclasses, 23% of all documents could not be linked to only one definite individual subclass (binary system or liver) due to the common features characterising these subclasses.

3. METHOD AND PROPOSED SYSTEM

The method proposed in this research is depicted in Figure 1. As shown, multiple steps are required to achieve the goals of this study. this system has four stages, first the system called preprocess step. it reseve unstructured data and applied the cleaning data tools on data. the second step was feature extraction. Then, therd step was feature selection. The last part was classification and evaluation.

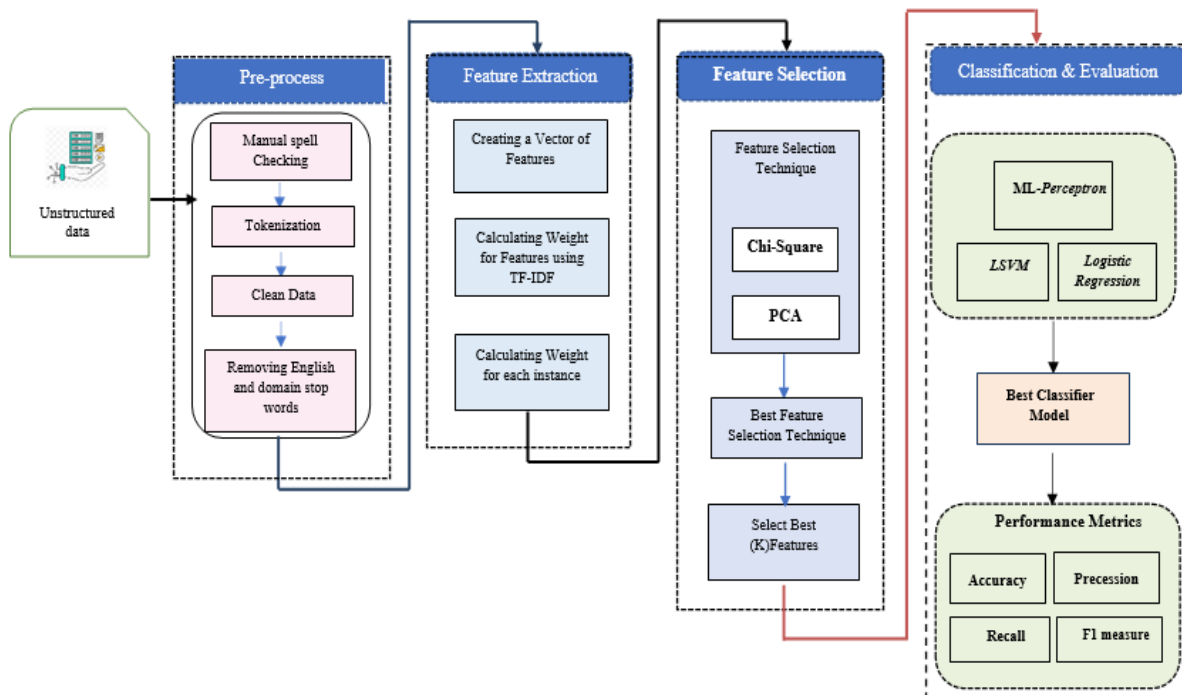


Figure 1. Illustration of the architecture of the proposed system

Dataset: the encounter dataset records interactions between a patient and healthcare provider(s) for the purpose of providing healthcare service(s) or assessing the health status of the patient. A patient encounter is further characterised by the setting in which it takes place, creating subsets of ambulatory, emergency, home health, inpatient, and virtual encounters. The date created was taken as 2018-09-20 and the date modified as 2019-11-01, with spatial coverage being the entire United States

The resulting data set contains 5,448 cases in 26 columns. Most of the data, however, is contained in the dscription column, free text natural language description section. This column is thus used for transcription and the assignation of medical specialty. This column was thus assumed to contain the medical speciality for each case, which was then used as a label.

3.1. Pre-processing steps

This stage was divided into two separate processes. The first was the pre-processing of structured data, while the second was the pre-processing of the remaining unstructured data to it suitable for text mining algorithms. For the unstructured data, these pre-processing steps included:

- Manual spellchecking: a spellchecker is a software tool that identifies words that may have been misspelled and suggests possible corrections. Every word from a text is compared to an appropriate lexicon, and when a word is not in the dictionary, it is highlighted as an error [14]. This step involves

making a sensitive comparison of each word with the closest matched words in dictionary where the result looks different according to the selected word. The dataset employed in this study includes clinical report-based multi-class classifications written by medical staff; it thus contained multiple misspelled words and transcription errors that np automated spellcheck package could handle. Manual spellchecking was thus used to return a set of possible candidates for each misspelled word, allowing the user to select the best matching word according to the scope of the reports.

- Tokenisation: in this study, tokenisation was performed by dividing the longer strings of medical reports into smaller pieces or tokens (words) based on spaces between them.
- Data cleaning: This was performed in two steps. i) Conversion to lower case: This step converted all characters to lowercase to simplify the NLP task. ii) Removing punctuation, symbols, and special characters: This step eliminated unnecessary information such as '!', '"', '#', '\$', '%', '&', '""', '(', ')', '*', '+', ',', '-', '.', ':', ';', '?', '[', '\\', ']', '_ ', '\'', '{', '|', '}', and '~' to render text in an appropriate form for NLP.
- Removing all English and domain stop words: this stage filtered out all words that contribute little to the overall meaning of a text such as “the”, “to”, and “this”. For this study, the library NLTK, in addition to a set of prespecified medical words (“mg”, “patient”, “tb”, “It”,...) was used to clean the data of stop words.

3.2. Feature extraction (FE)

The task of converting a particular text into a vector-based on space is an important part of text processing that can help extract the most important features from the text. FE methods are necessary to extract features as text data includes formats that are not accepted by machine learning techniques, and the features of such text must thus be extracted in a specific format that is accepted by such algorithms. In this research, the term frequency-inverse document frequency (TF-IDF) feature extraction method was used.

TF-IDF is a term derived from the fact that inverted document frequency can be used a numerical statistic that identifies the importance of a word in a set of documents [15]. Mathematically, TF-IDF is the result of the application two scales: TF and IDF. The TF-IDF for each term can thus be calculated using (1) [15].

$$TF - IDF (t, d) = TF (t, d) \times IDF (t) \quad (1)$$

where TF is defined as the number of times the term t occurs in document d , which is equivalent to the bags of words (BoW) approach. IDF thus refers to the statistical weight used to measure the significance of a term in a set of documents, which can be calculated using (2) [15].

$$IDF (t) = \log [(n)/(df (t))] \quad (2)$$

where the total number of documents in the document set is denoted by n , and $df(t)$ indicates the document frequency of t . Document frequency is thus the number of documents in the set of documents containing the term t . Once texts are converted using TF-IDF, and given appropriate weights, machine learning algorithms can then be used.

3.2.1. Creating a vector of features

Unstructured text is not suitable for use directly by machine learning techniques. As the pre-processing stage removes all punctuation, stop words, and other irrelevant features, the role of FE techniques is to transform a particular text into a matrix appropriate for machine learning algorithms, with rows representing the texts and columns representing the extracted features or words. to make it able to apply the feature selection. TF-IDF tools were used in this step.

3.2.2. Calculating weights for features

In this work, Term Frequency-TF-IDF was employed for transforming unstructured text into features based on term frequency and occurrence. This technique takes into account the number of times that a term appears in all documents and in the document set. Once the TF and IDF values are obtained according to (3), it is possible to calculate the TF-IDF:

$$TF - IDF_i = t_{fi} \times \log dN_{fi} \quad (3)$$

where t_{fi} = number of times, the term i appears in a document j , N refers to the total number of documents, and d_{fi} = number of documents that contain term i .

3.3. Feature selection (FS)

FS is a process that automatically selects those features in the data that contribute most to the prediction variable or output of interest. FS can thus be used in data pre-processing to achieve efficient data reduction. The best feature selection techniques for this work were identified as the Chi-square test and principal component analysis (PCA). The Chi-square test is a common statistical technique used for assessing categorical features in a dataset [16]. The Chi-square value between each feature and the target is calculated, and the desired number of features with the best Chi-square scores thus selected [17]. Principal component analysis (PCA) is a dimensionality reduction technique used to extract features from a dataset by reducing the dimensionality of the dataset based on employing matrix factorisation. This projects the dataset completely into a lower dimension while attempting to preserve the variance.

3.4. Data mining

Various AI techniques for text mining are used to automatically process data and generate useful or valuable insights, enabling users to make decisions based on the information extracted from such data. Text mining identifies facts, assertions, and relationships buried in blocks of textual big data that, without excavation, would never be discovered and which would remain buried indefinitely. Excavation can, however, extract the useful information, and then transform it into a structured model that can be analysed further or even presented directly without analysis. In terms of such data classification, several different techniques, such as random forest and logistic regression methods, have been proposed. Although many other data mining techniques have been utilised in the literature, only the methods most relevant to this research are presented here.

3.4.1. Support vector machine (SVM)

SVM is a supervised machine learning algorithm in which each data element is plotted as a point in an n-dimensional space representing n available features, with the value of each feature being a given coordinate value. The classification process is achieved by finding the hyper-level that most clearly distinguishes between the two categories. SVM was invented by Vapnik (1995) based on the theory of statistical learning, and it was originally designed for pattern recognition and multidimensional regression estimation, which can be used to solve linear or nonlinear problems [18]. SVM is only effective in high-dimensional spaces, though it remains effective in cases where the number of samples is less than the number of dimensions [14]. It is efficient in terms of memory use because it uses a subset of training points in the decision function (support vectors), and it is versatile in that it is possible to define a custom kernel as well as various kernel functions for the decision function. However, it is ineffective if the number of features is significantly greater than the number of samples, and it cannot provide probabilistic estimates directly [19].

3.4.2. Multiple layer perceptron (MLP)

MLP is an artificial neural network (ANN) with one or more hidden layers where the nodes in each layer are made up of nonlinearly activated neurons. MLP thus consists of one or more input layers, one or more hidden layers, and output layers where outputs from nodes are interconnected in a feed-forward direction [20]. MLP uses backpropagation (BP) technology for training that consists of redirection and back feed phases. The input is fed in the first stage to the nodes of the first hidden layer to perform activities related to the activation function passing from the input layer to the output layer, while in the second stage, the error between the desired and actual value is used to adjust the learning weights based on spreading the input layer from the output layer [21].

3.4.3. Logistic regression (LR)

LR is one of the most important statistical techniques in data mining and one of the most commonly used techniques by which statisticians and scientific analysts analyse and classify proportional and binary response data sets [22]. It is particularly popular for modelling binary data in health sciences and biostatistics, as it excels in determining the effects of most explanatory variables on a dichotomous outcome variable. LR is thus used to describe data and to explain the relationships between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables [23]. For maximum likelihood estimation, the asymptotic properties of the model parameters are usually used for statistical inference. However, logistic regression models are not suitable where cells with a value equal to zero occur in the contingency table, as this causes serious numerical problems [24]. The most important advantage of LR is its ability to provide probabilities naturally that extend to multi-class classification problems. Another advantage is that most of its optimisation is unrestricted, the same principles used in linear regression are followed in the analysis of linear regression models in most of the methods used in LR [25].

4. RESULTS AND DISCUSSION

The first Table 1 shows what happens when the model in Figure 1 receives data and applies a pre-processing stage to it, generating mining results by applying the three algorithms Lagrangian support vector machine (LSVM), ML-Perceptron, and logistic regression to calculate data quality over 10 cross validations. The next Table 2 shows the data quality of the same data when this is applied on the model in Figure 1 without the feature selection stage; a total of 10 cross validations were run. Table 3 shows the advantages of adding feature selection in terms of decreasing the training time for each algorithm, based on comparing the training times with and without feature selection and Figure 2 show the Performance Metrics of Three Algorithms with Chi-Square FS and Figure 3 without FS technique.

Table 1. Performance metric for the three algorithms with Chi-Square technique (10 cross validations)

Text Mining Algorithm	Accuracy Metrics			
	Accuracy	F1-Measure	precession	Recall
LSVM	98.49	96.86	98.68	95.39
ML-Perceptron	99.39	99.27	99.16	98.71
Logistic regression	97.67	94.38	98.65	91.78

Table 2. Performance metric for the three algorithms without FS technique (10 cross validations)

Text Mining Algorithm	Accuracy Metrics			
	Accuracy	F1-Measure	precession	Recall
LSVM	97.89	96.10	98.60	94.18
ML-Perceptron	98.79	98.82	98.62	97.90
Logistic regression	97.55	93.91	98.74	91.26

Table 3. Training time of the three algorithms

Text Mining Algorithm	Training Time (second)	
	Without FS technique	Without FS technique
LSVM	3.06	3.44
ML-Perceptron	15.08	15.5
Logistic regression	0.77	0.81

A consecutive series of 5000 medical reports from the electronic medical record (EMR) dataset was evaluated. NLP with TF-IDF and Chi-Square feature selection techniques was applied on this report to train the algorithm to classify items into ten medical groups. On reviewing the results obtained after applying AI techniques to unstructured data outside of the feature selection reference line, accuracy was found to be acceptable (above 98%) due to the use of pre-processing techniques, represented by the natural language processing steps. However, training and implementation took a long time due to the multiplicity of features, as shown in Table 2. A feature selection step was added to reduce the number of features used, retaining only the most effective features. The addition of this step significantly contributed to increasing the accuracy of the implementation of all the algorithms used.

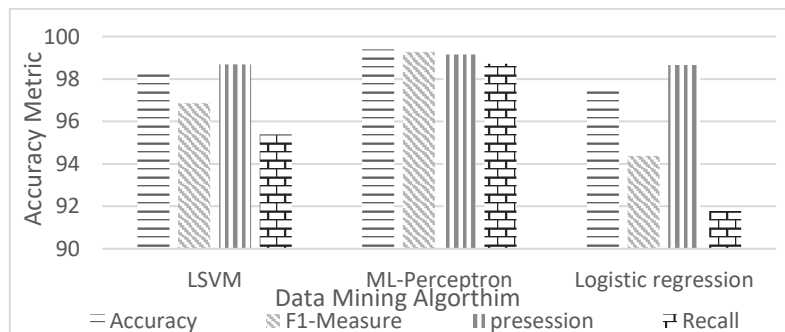


Figure 2. Performance metrics of the three algorithms with Chi-Square FS (10 cross validations)

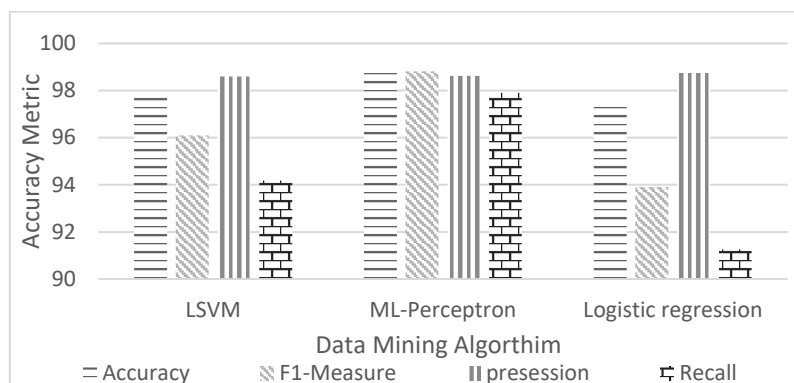


Figure 3. Performance metrics of the three algorithms without FS technique (10 cross validation)

The proposed model with TF-IDF was shown to be effective in identifying the best features and removing inappropriate or bad influences on unstructured data based on NLP techniques. In general, promising results emerged for all applied algorithms. The best accuracy achieved was with the ML-perceptron at 99.39; this also offered a reduction in the amount of time required for training where feature selection was used. Processing tools should be selected according to the characteristics of the data and the principles of dataset design followed. Even a design method which shows great performance in general contexts may suffer from performance variation in specific biomedical fields.

5. CONCLUSION.

From the results of this research, the following points can be concluded: i) Natural language processing (NLP) is the best way to determine a patient's medical classification and extract relevant data from electronic records, as this offers high reliability and accuracy, helping create better clinical databases. ii) A pre-processing step using NLP helps to make the initial unstructured data amenable to processing and mining and rids it of excess and useless attachments in order to allow greater benefit to be derived from the medical information within the initial records. iii) The addition of the selection feature leads to a noticeable increase in the accuracy of the results, especially when using the *Chi-Square* technique, which offers the best results. vi) The addition of the FS reduces the training time of the relevant algorithms, should speed up execution. ML-Perceptron was found to be the best algorithm in terms of accuracy. An algorithm can be trained using the database for the purpose of pre-classifying any medical report, even if the incoming reports were not originally classified based on the previous training values of the model.




REFERENCES

- [1] G. Chakraborty, M. Pagolu, and S. Garla, *PREVIEW: Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS*. SAS Institute, 2013.
- [2] M. C. R. Patel, M. V. K. Jadeja, and M. J. N. Solanki, "Review on natural language processing with artificial intelligence," *EPRA International Journal of Multidisciplinary Research (IJMR)*, vol. 6, no. 3, p. 143, 2020.
- [3] A. M. Nancy and R. Maheswari, "A review on unstructured data in medical data," *Journal of Critical Reviews*, vol. 7, no. 13, pp. 2202–2208, 2020.
- [4] M. Biniz, R. E. Ayachi, and M. Fakir, "Ontology matching using BabelNet dictionary and word sense disambiguation algorithms," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 5, no. 1, p. 196, Jan. 2017, doi: 10.11591/ijeecs.v5.i1.pp196-205.
- [5] M. Tayefi *et al.*, "Challenges and opportunities beyond structured data in analysis of electronic health records," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 13, no. 6, Nov. 2021, doi: 10.1002/wics.1549.
- [6] N. N. Alabid and Z. D. Katheeth, "Sentiment analysis of twitter posts related to the covid-19 vaccines," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 24, no. 3, pp. 1727–1734, 2021, doi: 10.11591/ijeecs.v24.i3.pp1727-1734.
- [7] A. K. Jha *et al.*, "Use of electronic health records in U.S. hospitals," *New England Journal of Medicine*, vol. 360, no. 16, pp. 1628–1638, Apr. 2009, doi: 10.1056/nejmsa0900592.
- [8] I. Li *et al.*, "Neural natural language processing for unstructured data in electronic health records: a review," *arXiv preprint arXiv:2107.02975*, p. 33, Jul. 2021, doi: <https://doi.org/10.48550/arXiv.2107.02975>.
- [9] N. L. I. Rusli, A. Amir, N. A. H. Zahri, and R. B. Ahmad, "Snake species identification by using natural language processing," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 13, no. 3, pp. 999–1006, Mar. 2019, doi: 10.11591/ijeecs.v13.i3.pp999-1006.
- [10] A. Esteva *et al.*, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24–29, Jan. 2019, doi: 10.1038/s41591-018-0316-z.
- [11] R. Cohen, I. Aviram, M. Elhadad, and N. Elhadad, "Correction: Redundancy-aware topic modeling for patient record notes," *PLoS ONE*, vol. 9, no. 11, p. e114677, Nov. 2014, doi: 10.1371/journal.pone.0114677.




- [12] J. Hammoud, A. Vatan, N. Dobrenko, N. Vedernikov, A. Shalyto, and N. Gusarova, "New Arabic Medical Dataset for Diseases Classification," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13113 LNCS, 2021, pp. 196–203.
- [13] M. Khachidze, M. Tsintsadze, and M. Archvadze, "Natural language processing based instrument for classification of free text medical records," *BioMed Research International*, vol. 2016, pp. 1–10, 2016, doi: 10.1155/2016/8313454.
- [14] N. Gupta and P. Mathur, "Spell checking techniques in NLP: a survey," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, no. 12, pp. 217–221, 2012.
- [15] M. Kannan, V. Gurusamy, S. Vijayarani, Ilamathi, and J. Nithya, "Preprocessing techniques for text mining preprocessing techniques for text mining," *International Journal of Computer Science & Communication Networks*, vol. 5, no. October 2014, pp. 7–16, 2015.
- [16] I. S. Thaseen and C. A. Kumar, "Intrusion detection model using fusion of chi-square feature selection and multi class SVM," *Journal of King Saud University - Computer and Information Sciences*, vol. 29, no. 4, pp. 462–472, Oct. 2017, doi: 10.1016/j.jksuci.2015.12.004.
- [17] A. J. Ferreira and M. A. T. Figueiredo, "Efficient feature selection filters for high-dimensional data," *Pattern Recognition Letters*, vol. 33, no. 13, pp. 1794–1804, Oct. 2012, doi: 10.1016/j.patrec.2012.05.019.
- [18] M. Awad and R. Khanna, "Support vector machines for classification," in *Efficient Learning Machines*, Berkeley, CA: Apress, 2015, pp. 39–66.
- [19] T. S. R. Sai, "Comparing the performance of various SVM classification techniques: a survey," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 13, pp. 1129–1136, 2021.
- [20] G. Xiao, J. Xing, and Y. Zhang, "Surface roughness prediction model of GH4169 superalloy abrasive belt grinding based on multilayer perceptron (MLP)," *Procedia Manufacturing*, vol. 54, pp. 269–273, 2020, doi: 10.1016/j.promfg.2021.07.042.
- [21] K.-L. Du and M. N. S. Swamy, "Multilayer perceptrons: Architecture and error backpropagation," in *Neural Networks and Statistical Learning*, London: Springer London, 2014, pp. 83–126.
- [22] A. Kouhpeima, S. Feyznia, H. Ahmadi, and A. R. Moghadamnia, "Landslide susceptibility mapping using logistic regression analysis in Latyan catchment," *Desert*, vol. 22, no. 1, pp. 85–95, 2017, doi: 10.22059/jdesert.2017.62181.
- [23] M. Kirişci, "Comparison of artificial neural network and logistic regression model for factors affecting birth weight," *SN Applied Sciences*, vol. 1, no. 4, p. 378, Apr. 2019, doi: 10.1007/s42452-019-0391-x.
- [24] A. Diop, A. Diop, and J. F. Dupuy, "Maximum likelihood estimation in the logistic regression model with a cure fraction," *Electronic Journal of Statistics*, vol. 5, no. none, pp. 460–483, Jan. 2011, doi: 10.1214/11-EJS616.
- [25] S. Sperandei, "Understanding logistic regression analysis," *Biochemia Medica*, vol. 24, no. 1, pp. 12–18, 2014, doi: 10.11613/BM.2014.003.

BIOGRAPHIES OF AUTHORS



Hasanen Abdul-Jawad Hussain Almuhan    Bachelor of Engineering Master's Student at University of Kerbala, work as associate Chief Engineer in ministry of interior I graduated from Babylon university, coolage of engineering, electrical department in 2006, know I will study Master degree in Kerbala university. Control and computer engineering. He can be contacted at email: hasanen.a@s.uokerbala.edu.iq.



Hawraa Hassan Abbas    Doctor of Engineering. Professor at University of Kerbala, a Ph.D. from CardiffUniversity/UK. She received herB.Sc. degree in computer engineeringfrom Baghdad University, Iraq andM.Sc. degree in computer engineering also from Baghdad University, Iraq. Her research interests include 3D face modeling, classification of facial traits, image processing, computer network design, genetic associations, computer vision. She can be contacted at email: Hawraa.h@uokerbala.edu.iq.