Clinical Text Classification with Word Representation Features and Machine Learning Algorithms

https://doi.org/10.3991/ijoe.v19i04.36099

Laiali Almazaydeh[1(✉)], Mohammed Abuhelaleh[2], Arar Al Tawil[3], Khaled Elleithy[4]
[1]Department of Software Engineering, Al-Hussein Bin Talal University, Ma'an, Jordan
[2]Department of Computer Science, Al-Hussein Bin Talal University, Ma'an, Jordan
[3]Abdul Aziz Al Ghurair School of Advanced Computing, Luminus Technical University College, Amman, Jordan
[4]Department of Computer Science and Engineering, University of Bridgeport, Bridgeport, USA
laiali.almazaydeh@ahu.edu.jo

Abstract—Clinical text classification of electronic medical records is a challenging task. Existing electronic records suffer from irrelevant text, misspellings, semantic ambiguity, and abbreviations. The approach reported in this paper elaborates on machine learning techniques to develop an intelligent framework for classification of the medical transcription dataset. The proposed approach is based on four main phases: the text preprocessing phase, word representation phase, features reduction phase and classification phase. We have used four machine learning algorithms, support vector machines, naïve bayes, logistic regression and k-nearest neighbors in combination with different word representation models. We have applied the four algorithms to the bag of words, to TF-IDF, to word2vec. Experimental results were evaluated based on precision, recall, accuracy and F1 score. The best results were obtained with the combination of the k-NN classifier, and the word represented by Word2vec achieving an accuracy of 92% to correctly classify the medical specialties based on the transcription text.

Keywords—clinical text, classification, logistic regression, support vector machines, k-nearest neighbors, naïve bayes, word of bag, TF-IDF, word2vec

1 Introduction

Natural Language Processing (NLP) is a process utilized for analyzing programming languages to help develop tools and techniques that can allow computers to perform various tasks based on the knowledge of humans. It has gained widespread attention due to its numerous applications in various fields like machine translation, email spamming detection, medical inquiries, and question answering. Currently, NLP field consists variety of cognitive models, linguistic theories, and engineering issues; and its applications can be classified into different classifications. One of them is to classify it into: speech recognition, natural language interface, story understanding, text generations, discourse Management, and text classification [1–2].

Text classification is an important task in NLP which is a process of extracting structured and unstructured information from a text and then classify it according to

some rules. One of the most important fields of text classification is the clinical text classification, where a text is extracted from the Electronic Medical Records (EMR) and classified into different classifications. The clinical text classification process involves different tasks that may differ from system to another. For example, clinical text classification may provide smoking-status detection, patient classifications for different purposes, and classification of obesity [3–4]. EMR are widely adapted in medical providers' systems, which led to creating large containers of structured and unstructured clinical data. These data can be classified to be leveraged in biomedical research and in health care delivering. On the other side, Text classification of EMR involves specific challenges compared to other related systems. These challenges include: the misbalancing of the dataset, misspellings, semantic ambiguity, and abbreviations [5].

Recently, a clinical text classification widely adopts machine learning to demonstrate information extraction and text classification. These led the community of NLP to give more attention of applying machine learning and improving the performance of text classification. Usually, the most common first stage is to extract some features, such as bag of n-grams and bag of words [6].

In this work, we analyze the impact of various word representations such as BOW, TF-IDF and Word2vec and the impact of various machine learning algorithms such as Support Vector Machine (SVM), Naïve Bayes, k-NN and Logistic Regression (LR) on the performance of clinical text classification tasks. The experiment will be conducted on the medical literature dataset to investigate which algorithms has the most efficiency towards clinical text classification performance.

This paper is organized as follows: Section 2 offers the related research. Section 3 describes the details of our proposed methodology. Section 4 demonstrates the experimental results and evaluation. Section 5 finally outlines the conclusion and future works.

2 Related works

In this section, a number of recent studies, which are based on different machine learning techniques for clinical text classification, are reviewed in general.

Yumeng Guo et al. [7] proposed an algorithm for multi-label text classification of clinical records by grouping the selection of embedded feature of those records. They claimed that this method should produce a firm feature classification and lessen the negative effects of educating dataset. In addition, they claimed that this method may gain more performance if it is adjusted to multi-label data structure for clinical text data. The algorithm employed the forward search strategy and prediction risk in its work to assess the importance of features at the time the feature subset is generated to enhance the classifier's performance. The algorithm mainly depends on the classifier's learning capability and the measure of employed evaluation for calculating prediction risk. The authors applied their algorithm to 1566 clinical records marked by disease codes from a dataset that was created by the Computational Medicine Center (NLP Challenge). The records' contents mainly contained reports from some radiologists regarding some patients written in the form of free text. The experiment extracted the bag of word features from that raw text and then transformed the counts of words into TF-IDF features. After filtering of stop words and stemming of word, the algorithm only kept the word of frequencies of the highest 232 words. Then a batch of ICD-9-CM codes

are produced to represent disease labels. This batch consists of carefully entries of classified disease which are labeled by distinguished numbers. These numbers can be used to categorize the clinical records into their corresponding diseases.

Mei-Sing Ong and others [8] proposed an automated categorization algorithm of clinical incident reports by using the classification of statistical text. They developed three classifiers of statistical text to identify two clinical incidents' classes: incorrect patient identification and inadequate clinical handover. They developed their classifiers based on the SVM with radial-basis function and with linear Kernel, and Naïve Bayes algorithm. These algorithms were trained and tested on some incident reports that are submitted by some public hospitals. The target was to identify the two clinical incidents clarified before. They trained each classifier on 600 reports and tested on 372 reports. Then, they evaluate the results using some standard measures of accuracy such as recall, precision, area under the curve, and F-measure. They used a limited number of incidents from the AIMS dataset for developing and testing their algorithm. The results they performed show good performance on clinical handover categorization and on the identification of the patient's incidents. Their results showed that the algorithm they provided should perform well in categorizing identification incidents of patient and on classifying clinical handover. They achieved an accuracy around 80% with most classifiers using a small sample of 100 records.

Yijun Shao et al. [6] made a comparison between two of the most common used techniques for clinical text classifications (i.e., Word Embedding Features and Bag-of-Words Features). They applied doc2vec and word2vec features on some set of clinical text classification functions and then they compared the results using the features of normal bag-of-words (BOW). They generated their dataset from Veterans Affairs (VA) electronic records which are stored in VINCI database. The generated dataset was used to study the use of alternative and complementary Medicine (called CAM) through veterans. They also annotated their dataset by using a random number specimen of clinical notes taken from VINCI. At least, one predefined keyword should be contained in each note. They also divided their dataset into six modalities: Biofeedback, Meditation, Acupuncture, Guided Imagery, Yoga, and Tai-chai. For each modality, they selected a small subset for human annotation. These subsets are then used for developing some extraction tools of natural language processing. All these subsets were labeled using meaningful labels and the annotation of the original categories was labeled using binary labels. For classification and learning, linear kernel SVM was used. Their study showed that word2vec features worked better than the features of BOW-1-gram. Moreover, adding 2-grams to BOW showed mixed results.

Li Qing and others [9] proposed a novel algorithm for text classification based on neural network. In their proposal, the features were extracted from the Bidirectional gated recurrent unit (BIGRU) using convolutional layer to access the succeeding and the preceding sentence features. They employed attention mechanism to gain the sentence representation with the word weights according to their importance. BIGRU was used to encode the obtained sentences from sentence representation and then decoded it using an attention mechanism to retrieve a document representation with the weight of importance. The last stage of their algorithm is to obtain a medical text category using a classifier. They run their experiments on 4 medical datasets which included 2 medical literature datasets, and 2 medical record datasets. They evaluated their algorithm on the task of medical text classification using two datasets: Traditional Chinese

medicine (TCM), and an open access dataset from CCKS conference. Their results showed effective performance on text classification.

Kevin Chai et al. [10] Examined the feasibility of using statistical text classification to identify health information technology incidents. They applied their tests on a database belonging to USA Food and Drug Administration (FDA) and to Manufacturer and User Facility Device Experience (MAAUDE). A sample of 570272 incidents, including 1524 reported HIT incidents, was used in their test. They evaluated normalized logistic regression on balanced and stratified datasets for validation, testing, and training. They also performed dataset's feature extraction, cross-validation, error analysis, preparation, feature selection, performance evaluation, and classification. In addition, they examined some techniques of feature-selection like stemming, principal component analysis, removing stop words and short words, and lemmatization. Their results showed that stemming was performing better than other techniques. Moreover, they reduced the size of the feature set to 79% to achieve an improvement of recall to 0.989 but reducing the precision to 0.165. They concluded that the statistical text classification is a feasible technique to identify HIT report through a large-size incidents' database. On other side, automated identification is expected to increase the percentage for detection, analysis, and addressing of HIT problems. They also advised to apply some supervised learning and more investigating when dealing with such big-data analysis for patient safety incidents.

3 Materials and methods

This section discusses the methodology as indicated in Figure 1 through the steps in a clinical text classification pipeline such as text pre-processing, word representation, features reduction, and classification.

3.1 Dataset preparation

Finding medical data is very difficult due to the HIPAA privacy policy, nevertheless, few medical record datasets and medical literature datasets are available for the related research. These are: TCM [11], Hallmarks [12], and AIM [12]. But these datasets have some shortcomings such as limited medical categories reaching only five classes and small dataset size. In our work, we used medical transcription dataset from

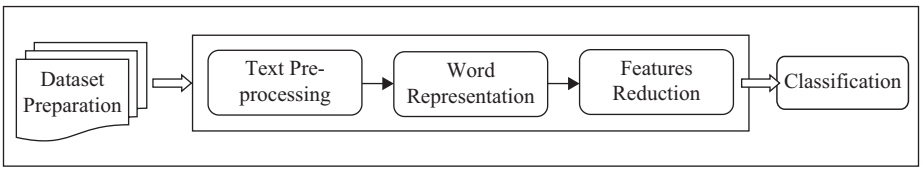

Fig. 1. Clinical text classification pipeline

mtsamples.com which has sampled transcriptions belonging to different medical specialties. In this dataset there are around 40 categories, 140214 sentences in their transcription and 35822 unique words in the transcription.

The number of transcription records are quite less for some of the categories, so we filtered out these categories to keep which have at least more than 50 samples, then the number of categories was reduced from 40 to 20 with 4324 records as presented in Table 1. Figure 2 shows the word cloud of these filtered medical categories. Table 2 shows different examples of medical transcriptions which contain lots of vocabularies and abbreviations, every given transcription indicates specific medical specialty.

Table 1. Medical specialty records

Medical Specialty	Samples
Cardiovascular/Pulmonary	371
Consult – History and Phy	516
Discharge Summary	108
ENT – Otolaryngology	96
Emergency Room Reports	75
Gastroenterology	224
General Medicine	259
Hematology – Oncology	90
Nephrology	81
Neurology	223
Neurosurgery	94
Obstetrics/Gynecology	155
Ophthalmology	83
Orthopedic	355
Pain Management	61
Pediatrics – Neonatal	70
Psychiatry/Psychology	53
SOAP/Chart/Progress Notes	166
Surgery	1088
Urology	156

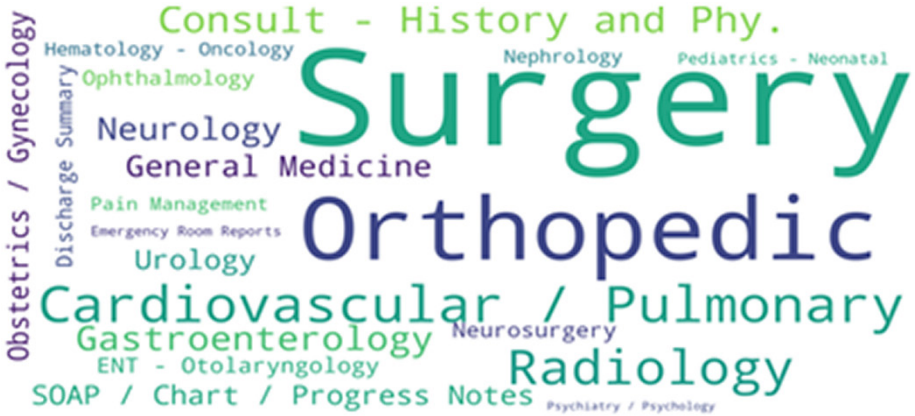

Fig. 2. Word cloud of the clinical specialty

Table 2. Different examples of medical transcriptions with their corresponding medical specialty

Medical Specialty	Transcription
Urology	This is a 66-year-old male with signs and symptoms of benign prostatic hypertrophy, who has had recurrent urinary retention since his kidney transplant. He passed his fill and pull study and was thought to self-catheterize in the event that he does incur urinary retention again.
Rheumatology	A 71-year-old female who I am seeing for the first time. She has a history of rheumatoid arthritis for the last 6 years. She is not on DMARD, but as she recently had a surgery followed by a probable infection.
Surgery	Cholecystitis and cholelithiasis. Laparoscopic cholecystectomy and intraoperative cholangiogram. The patient received 1 gm of IV Ancef intravenously (intravenous) piggyback. The abdomen was prepared and draped in routine sterile fashion.

3.2 Text Pre-processing

The text pre-processing phase comprises of the following steps: Data cleaning, stop words removal, lemmatization, and tokenization. Table 3 shows the effects of these steps on part of a sample input text (record no. 500 with "Pediatrics" medical specialty).

Table 3. The effect of pre-processing steps on a sample input text

Pre-Processing Steps	Effect of Pre-Processing Steps on a Sample Input Text
Input text	SUBJECTIVE: This is a 12-year-old male who comes in for healthy (health) checkups and sports physical. No major concerns today. He is little bit congested at times. He has been told he is allergic to grasses. They have done over-the-counter Claritin and that seems to help but he is always sniffling mother reports. He has also got some dryness on his face as far as the skin and was wondering what cream he could put on.
Data Cleaning	SUBJECTIVE This is a year old male who comes in for healthy checkups and sports physical No major concerns today He is little bit congested at times He has been told he is allergic to grasses They have done over the counter Claritin and that seems to help but he is always sniffling mother reports He has also got some dryness on his face as far as the skin and was wondering what cream he could put on
Text Lowering	subjective this is a year old male who comes in for healthy checkups and sports physical no major concerns today he is little bit congested at times he has been told he is allergic to grasses they have done over the counter Claritin (clarity) and that seems to help but he is always sniffling mother reports he has also got some dryness on his face as far as the skin and was wondering what cream he could put on
Lemmatization	subjective year old male come healthy checkup sport physical major concern today little bit congested time told allergic grass done over the counter claritin seems help always sniffling mother report also got dryness face far skin wondering cream could put on
Tokenization	['subjective', 'year old', 'male', 'come', 'healthy', 'checkup', 'sport', 'physical', 'major', 'concern', 'today', 'little', 'bit', 'congested', 'time', 'told', 'allergic', 'grass', 'done', 'over the counter', 'claritin', 'seems', 'help', 'always', 'sniffling', 'mother', 'report', 'also', 'got', 'dryness', 'face', 'far', 'skin', 'wondering', 'cream', 'could', 'put', 'on']

Data cleaning. In this step a set of functions are defined to clean the transcription data, using these functions most of the punctuation, special characters, digits, URLs, stop words, and irrelevant text are removed [13].

Text lowering. In this step each of the capital characters are converted into their corresponding small characters.

Lemmatization. In this step the morphological analysis is taken into consideration to reduce the variability in the words, so the words which have similar meaning will be mapped to the same word [14]. What distinguishes lemmatization is that it attempts to identify the correct lemma depending on the context.

Tokenization. Tokenization is the process of splitting text into smaller pieces and each piece is then called a token. The most common token size is a word [15].

3.3 Word representation

The word representation phase includes converting text data into some vector representation so that the algorithms will be automatically able to understand analogies and generalize that word. Word representation models varies between classical models and representation learning models [16]. In our work, three common models have been adopted for clinical text classification task. These models are discussed below.

Bag-of-Words (BOW). In learning from text, each individual word can't be used as an input feature, because then long document would require different input space than short document. Instead, BOW approach is used to count the frequency of words, then each word will be mapped into a frequency count [17].

Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF was presented by [18] as a weighting factor for feature extraction. It is a numerical statistic that is intended to reflect how important a word to a document in a collection or corpus. The TF is the number of times a term appears in a document and the IDF is the number of times of documents that contains a particular word.

Word2vec. Word2vec was developed by Tomas Mikolov and his team at Google in 2013 [19]. It uses a neural network architecture which tune its weights using backpropagation and gradient descent to convert words into vectors. There are two learning models to produce words representation, one called the continuous bag-of-words (CBOW) and the other one called the continuous skip-gram model. The difference between these models that, the CBOW model uses the context words to predict the target word and the continuous skip-gram model does the opposite, it uses the target word to predict the context of words [20]. So that words are embedded in vector space alongside related words based on their contextual meaning.

3.4 Features reduction

Principal component analysis (PCA): is one of the most important dimensionality reduction techniques. It takes a high dimensional space and applies various transformations onto it to get it to a lower dimensional space such that this lower dimensional space still captures as much of the dynamics in the original space [21].

The entire process of PCA is basically built to reduce input variable redundancy by creating a new set of variables where the variance along each subsequent variable is maximized.

3.5 Classification

After performing text preprocessing, word representation and features reduction, the classifiers should be implemented. We had considered four supervised classifiers namely logistic regression, support vector machines, k-nearest neighbors and naïve bayes to determine which has the best results. The mtsamples dataset was split into 75% for the training set and 25% for the testing sett. After splitting, pipeline had been used for implementing the classifiers. Pipelining is used for better flow of an algorithm.

The different classifiers which have been adopted for clinical text classification task are discussed below.

Logistic Regression (LR) Classifier. Logistic regression models the probability associated with the level of the response variable by finding the relationship between predicting variables and link function with this probability. Multinomial is the type of LR which is used to assess the relationship when the number of categories is three or more, and characteristics are not as per at the natural ordering of the levels [22–23], hence this type is used in the employed technique.

Support Vector Machine (SVM) Classifier. Support vector machine is a supervised learning algorithm for classification. Each object to be classified is represented as a point in an n-dimensional space and the coordinates of this point are usually called features. SVM performs the classification by drawing a hyper-plane whereas all points of one category are on one side of the hyper-plane and all points of the other category are on the other side. It could be there multiple such hyper-planes whereas SVM tries to find the one that best separates the two categories in the sense that it maximizes the distance to points in either category. This distance is called the margin and the points that fall exactly on the margin are called the supporting vectors [24].

Classification of text can be performed using linear kernel, as this kind of classification can be linearly separated, and linear kernel is a good choice when dealing with large sparse data vectors, hence this kernel is used in the employed technique.

k-nearest neighbors (k-NN) Classifier. k-NN is one of the supervised machine learning algorithms mostly used for classification. It classifies a data point based on how its neighbors are classified. Therefore, k in k-NN is a parameter that refers to the number of nearest neighbors to include in the majority of the voting process. i.e., a data point is classified by majority of votes from (of) its k nearest neighbors. Choosing the correct value of k is a process called parameter tuning so that there is no a significant bias in one direction or the other direction. Hence, this results in better accuracy [25].

Naïve bayes classifier. Naïve bayes is a probabilistic classifier which learns the probability of an object with certain features belonging to a particular group in class. The naïve bayes algorithm is called naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features.

The basis of the naïve bayes algorithm is the Bayes theorem which is alternatively known as Bayes rule or Bayes law. It gives a method to calculate the conditional probability that is the probability of an event based on previous knowledge available on the events [26].

4 Results

In this study, four classification performance measures were adopted, namely: accuracy, precision, recall and F1 score [27]. These measures are based on four possible outcomes, true positive (*TP*), false positive (*FP*), true negative (*TN*), and false negative (*FN*).

The classification accuracy refers to the ratio of correct decisions (i.e., true positive plus true negative) to the total number of cases. The equations of four classification performance measures are listed below:

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{1}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{2}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{3}$$

$$F1\ Score = 2 * \frac{(Recall * Precision)}{(Recall + Precision)} \tag{4}$$

The performance of the different classification algorithms and different word representation models is summarized in Table 4. Based on the results which are tabulated in Table 4, we find that the lowest accuracy is obtained when BOW was used, and the highest accuracy was obtained with the combination of the k-NN classifier, and the word represented by Word2vec achieving 92%.

These results can be attributed to the limited capability of BOW model to represent words only as a sparse matrix that has a lot of zero values, while Word2vec model able to retain most of the contextual information or linguistic information.

Table 4. Performance of the different classifiers and word representation models

Classifier Model	Word Representation	Precision	Recall	F1 Score	Accuracy
LR	**BOW**	0.81	0.80	0.79	0.83
	TF-IDF	0.88	0.89	0.88	0.89
	Word2vec	0.90	0.91	0.90	0.90
SVM	**BOW**	0.75	0.78	0.76	0.77
	TF-IDF	0.84	0.81	0.82	0.81
	Word2vec	0.88	0.85	0.86	0.85
k-NN	**BOW**	0.80	0.81	0.80	0.80
	TF-IDF	0.85	0.84	0.84	0.84
	Word2vec	0.91	0.89	0.90	0.92
Naïve Bayes	**BOW**	0.70	0.75	0.72	0.73
	TF-IDF	0.75	0.75	0.75	0.75
	Word2vec	0.88	0.86	0.87	0.88

5 Conclusion and future works

Word representation is a fundamental step in the process of machine learning for analyzing data. In this paper, three common word representation models were experimented: BOW, TF-IDF, and Word2vec for allocating words (word) vectors of the medical transcription dataset. These word representation models are used as features to train four machine learning classifiers: LR, SVM, k-NN, and Naïve Bayes to build clinical text classification model. Experimental results showed that a model built on Word2vec based k-NN classifier has achieved a highest (higher) result with (in) an average of accuracy 92%. These good results can be attributed to the capability to retain linguistic information by Word2vec model. The proposed approach represents a promising tool in clinical text classification challenging field that has a lot of beneficial applications such as disease classification for various purposes. Therefore, as future work, the experimental experiments will be validated on Covid-19 dataset if available, also deep learning techniques will be investigated as it promises more performance than which result from traditional techniques.

Funding Statement: This work was fully funded by the Deanship of Scientific Research and Graduate Studies at Al-Hussein Bin Talal University in Jordan under grant No. (31/2022).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

6 References

[1] M. Agarwal and A. Saxena, "An overview of natural language processing," International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 7, no. 5, pp. 2811–2813, 2019. https://doi.org/10.22214/ijraset.2019.5462

[2] I. Zeroual and A. Lakhouaja, "Data science in light of natural language processing: An overview," Procedia Computer Science, vol. 127, pp. 82–91, 2018. https://doi.org/10.1016/j.procs.2018.01.101

[3] H. Zhu and L. Lei, "The research trends of text classification studies (2000–2020): a bibliometric analysis," SAGE Open, vol. 12, no. 2, pp. 1–16, 2022. https://doi.org/10.1177/21582440221089963

[4] P. Thangaraj, B. Kummer, T. Lorberbaum, M. Elkind and N. Tatonetti, "Comparative analysis, applications, and interpretation of electronic health record-based stroke phenotyping methods," BioData Mining, vol. 13, no. 21, pp. 1–14, 2020. https://doi.org/10.1186/s13040-020-00230-x

[5] V. Garla, C. Taylor and C. Brandt, "Semi-supervised clinical text classification with Laplacian SVMs: An application to cancer case management," Journal of Biomedical Informatics, vol. 46, no, 5, pp. 869–875, 2013. https://doi.org/10.1016/j.jbi.2013.06.014

[6] Y. Shao, S. Taylor, N. Marshall, C. Morioka and Q. Zeng-Treitler, "Clinical text classification with word embedding features vs. Bag-of-Words features," in Proc. IEEE International Conference on Big Data, Seattle, WA, USA, pp. 2874–2878, 2018. https://doi.org/10.1109/BigData.2018.8622345

[7] Y. Guo, F. Chung and G. Li, "An ensemble embedded feature selection method for multi-label clinical text classification," in Proc. 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Shenzhen, China, pp. 823–826, 2016.

[8] M-S. Ong, F. Magrabi and E. Coiera, "Automated categorization of clinical incident reports using statistical text classification," Quality and Safety in Health Care, vol. 19, no. 6, pp. 1–7, 2010. https://doi.org/10.1136/qshc.2009.036657

[9] L. Qing, W. Linhong and D. Xuehai, "A novel neural network-based method for medical text classification," Future Internet, vol. 11, no. 12, pp. 1–13, 2019. https://doi.org/10.3390/fi11120255

[10] K. Chai, S. Anthony, E. Coiera and F. Magrabi, "Using statistical text classification to identify health information technology incidents," Journal of the American Medical Informatics Association, vol. 20, no. 5, pp. 980–985, 2013. https://doi.org/10.1136/amiajnl-2012-001409

[11] L. Yao, Y. Zhang, B. Wei, Z. Li and X. Huang, "Traditional Chinese medicine clinical records classification using knowledge-powered document embedding," in Proc. 2016 IEEE International Conference on Bioinformatics and Biomedicine, Shenzhen, China, pp. 1926–1928, 2016.

[12] S. Baker, I. Silins, Y. Guo, I. Ali, J. Hogberg et al., "Automatic semantic classification of scientific literature according to the hallmarks of cancer," Bioinformatics, vol. 32, no. 3, pp. 432–440, 2016. https://doi.org/10.1093/bioinformatics/btv585

[13] F. Iqbal, J. Hashmi, B. Fung, R. Batool, A. Khattak et al., "A hybrid framework for sentiment analysis using genetic algorithm-based feature reduction," IEEE Access, vol. 7, pp. 14637–14652, 2019. https://doi.org/10.1109/ACCESS.2019.2892852

[14] M. Karamibekr and A. Ghorbani, "Sentiment analysis of social issues," in Proc. 2012 International Conference on Social Informatics, DC, USA, pp. 215–221, 2012. https://doi.org/10.1109/SocialInformatics.2012.49

[15] E. Alawneh, M. AlFawa'reh, M. Jafar and M. AlFayoumi, "Sentiment analysis-based sexual harassment detection using machine learning techniques," in Proc. 2021 International Symposium on Electronics and Smart Devices (ISESD), Bandung, Indonesia, pp. 1–6, 2021. https://doi.org/10.1109/ISESD53023.2021.9501725

[16] U. Naseem, I. Razzak, K. Khan and M. Prasad, "A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models," Transactions on Asian and Low-Resource Language Information Processing, vol. 20, no. 5, pp. 1–35, 2021. https://doi.org/10.1145/3434237

[17] Y. Zhang, J. Rong and Z. Zhi-Hua, "Understanding bag-of-words model: A statistical framework," International Journal of Machine Learning and Cybernetics, vol. 1, no. 1, pp. 43–52, 2010. https://doi.org/10.1007/s13042-010-0001-0

[18] K. Jones, "Document retrieval systems," Chapter A Statistical Interpretation of Term Specificity and Its Application in Retrieval, Taylor Graham Publishing, London, UK, pp. 123–142, 1988.

[19] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient estimation of word representations in vector space," in Proc. 1st International Conference on Learning Representations (ICLR 2013), Arizona, USA, 2013.

[20] L. Wensen, C. Zewen, W. Jun and W. Xiaoyi, "Short text classification based on Wikipedia and Word2vec," in Proc. 2nd IEEE International Conference on Computer and Communications (ICCC), pp. 1195–1200, 2016. https://doi.org/10.1109/CompComm.2016.7924894

[21] S. Karamizadeh, S. Abdullah, A. Manaf, M. Zamani and A. Hooman, "An overview of principal component analysis," Journal of Signal and Information Processing, vol. 4, pp. 173–175, 2013. https://doi.org/10.4236/jsip.2013.43B031

[22] C-Y. Peng, K. Lee and G. Ingersoll, "An introduction to logistic regression analysis and reporting," The Journal of Educational Research, vol. 96, no. 1, pp. 3–14, 2002. https://doi.org/10.1080/00220670209598786

[23] A. Genkin, D. Lewis and D. Madigan, "Large-scale bayesian logistic regression for text categorization," Technometrics, vol. 49, no. 3, pp. 291–304, 2007. https://doi.org/10.1198/004017007000000245

[24] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, pp. 273–297, 1995. https://doi.org/10.1007/BF00994018

[25] H. Elzeheiry, S. Barakat and A. Rezk, "Different scales of medical data classification based on machine learning techniques: A comparative study," Applied Sciences, vol. 12, no. 2, pp. 1–20, 2022. https://doi.org/10.3390/app12020919

[26] M. Islam, Q. Wu, M. Ahmadi and M. Sid-Ahmed, "Investigating the performance of naïve bayes and k-nearest neighbor classifiers," in Proc. 2007 International Conference on Convergence Information Technology (ICCIT 2007), Gwangju, South Korea, pp. 1541–1546, 2007. https://doi.org/10.1109/ICCIT.2007.148

[27] L. Almazaydeh, S. Atiewi, A. Al Tawil and K. Elleithy, "Arabic music genre classification using deep convolutional neural networks (CNNs)," Computers, Materials and Continua, vol. 72, no. 3, pp. 5443–5458, 2022. https://doi.org/10.32604/cmc.2022.025526

7 Authors

Laiali Almazaydeh received her Ph.D. in Computer Science and Engineering in 2013 from University of Bridgeport, USA. She is a professor and the Dean of Faculty of Information Technology at Al-Hussein Bin Talal University, Jordan. Laiali has published more than 55 research papers in various international journals and conferences proceedings, her research interests include human computer interaction, pattern recognition, and computer security. She received best paper awards in 3 conferences, ASEE2012, ASEE2013 and ICUMT 2016. She can be contacted at email: laiali.almazaydeh@ahu.edu.jo.

Mohammed Abuhelaleh received his Ph.D. in Computer Science and Engineering from University of Bridgeport, USA in 2011. He is an associate professor and in September 2017 he was appointed the Dean of Faculty of Information Technology for two years at Al-Hussein Bin Talal University, Jordan. His research interests include object oriented, wireless sensor networks, and pattern recognition. He has published several conferences papers, and journals papers in these topics. He can be contacted at email: mabuhela@ahu.edu.jo.

Arar Al Tawil received his MSc. degree in 2021 from Jordan University, Jordan. He is a BTEC Lecturer in Machine learning, Data mining, and database development at Luminus Technical University College, Amman, Jordan. His research interests include Virtual reality, Augmented reality environments, and machine learning and data analysis. He can be contacted at email: arartawil@gmail.com.

Khaled Elleithy is the Associate Vice President for Graduate Studies and Research at the University of Bridgeport, Connecticut, USA. He is a professor of Computer Science and Engineering. His research interests include wireless sensor networks, mobile communications, network security, and quantum computing. He has published more than four hundred papers in national/international journals and conferences in his areas of expertise. He can be contacted at email: elleithy@bridgeport.edu.

Article submitted 2022-10-15. Resubmitted 2023-02-05. Final acceptance 2023-02-07. Final version published as submitted by the authors.