Proposed Model for Text Classification Using Natural Language Processing Techniques.

[1]Atef Ghalwash,[2] Esam Abu Elaseam & [3]Omnia Mamdouh

[1]Faculty of computers and artificial intelligence, Helwan university

[2,3] Faculty of commerce and business administration Helwan university, Cairo, Egypt

Abstract

The pathology reports contain unorganized text, and it can be difficult to access or search for clinical information within them. Automated text classification is an important method that can help manage and process a large number of digital documents, which are continuously increasing, by categorizing them into predefined classes. This process plays an important role in extracting and summarizing information, retrieving text, and answering questions. This research focused on using the LSTM, CNN, and Logistic Regression algorithms to classify pathology reports. The dataset used contained research papers with more than 6 pages, specifically cancer documents categorized into Thyroid, Colon, and Lung cancers. The results showed that the model had acceptable accuracy levels, with the LSTM algorithm achieving a high accuracy rate of 99%.

Keywords: Text Classification; Natural Language Processing; Long short-term memory; Convolutional Neural Network.

1 Introduction

In recent years, there has been a significant increase in the amount of digital text documents available. As a result, it has become crucial to efficiently categorize and organize these documents. Text classification refers to the process of grouping textual information based on its content [1]. The primary objective of text classification is to divide unstructured documents into respective categories based on their content [2]. This process finds application in various fields such as text summarization, information extraction, information retrieval, question answering, and sentiment analysis due to the widespread availability of textual information [1].

Text classification is used in different fields such as spam filtering, email routing, topic tracking, sentiment analysis, and web page classification. However, medical specialization is considered one of the most crucial areas where text classification plays an essential role.

Pathologists use unstructured and semi-structured pathology reports to record detailed observations of cells, organs, and tissue specimens. These reports contain a vast amount of information that is essential for advancing cancer research in areas such as treatment selection, case identification, prognostication, surveillance, clinical trial screening, risk stratification, retrospective study, and many others [3].

Retrieving crucial descriptive observations from pathology reports presents a significant challenge because a considerable part of the diagnosis is written in an unstructured, free-text format. State or national cancer registries responsible for tracking thousands or even millions of patients must rely on human experts to decode and convert this relevant information into a structured and normalized form. This process of manually coding information is expensive, time-consuming, prone to errors, and places inherent restrictions on the amount and variety of data that can be extracted [4].

The cancer research communities have recently turned their attention towards NLP as a solution to overcome the constraints of manual information extraction from pathology reports [5]. This study aims to create a Text classification model that uses Logistic Regression and LSTM techniques to automate the IE process on pathology reports. The primary objective is to develop an efficient and automated approach to extract information from these reports.

The paper is organized as follows: Section 2 presents a brief review of previous research on text classification and offers examples of text classifiers that have been developed using various techniques. In Section 3, the theoretical concepts and principles underlying the proposed text classifier are discussed in detail. Section 4 gives a detailed description of how the classifier works, including its pseudocode and algorithms for implementation. Time complexity of the classifier is examined, and results obtained using the classifier are analyzed in Section 5. Finally, Section 6 concludes the paper with a summary of the work and discusses potential future directions for research.

.

2 Literature review

The objective of this research is to investigate how well different deep learning algorithms perform in classifying medical notes with imbalanced disease classes. The study utilized seven AI models, including a CNN, Transformer encoder, and pre-trained BERT, as well as four traditional sequence neural network models - RNN, GRU, LSTM, and Bi-LSTM. These models were used to determine the presence or absence of 16 specific diseases based on discharge summary notes from patients. The best performing model was MLearn-ATC, achieving an accuracy of 0.93. However, it should be noted that this technique was not evaluated using a larger dataset or non-binary classification problems [6].

This paper explores the efficacy of a semi-supervised approach for Telugu news articles. The study uses semi-supervised clustering for pattern classification after initial categorization. Its goal is to evaluate the impact of n-gram feature selection on text classification for news articles using semi-supervised learning techniques. Results indicate that these approaches significantly boost performance, with Support Vector Machine achieving a 93.64% classification rate [7].

The authors suggest using pre-trained language models and logic rules to create prompts with sub-prompts for multi-class text classification. This method, called PTR, allows for encoding prior knowledge of each class into the prompt tuning. The experiments conducted on relation classification, a complex multi-class classification task, demonstrate that PTR outperforms existing baselines consistently. These results indicate that PTR has promise as an approach that utilizes both human prior knowledge and PLMs for challenging classification tasks, achieving an F1 score of 91.9 [8].

In this study, an intelligent approach that utilizes natural language processing technology (NLP) is proposed. The approach consists of three stages. Firstly, a Convolution Neural Network (CNN) based text classification model is developed to classify construction on-site reports by analyzing and extracting the features of report text. In the second stage, an improved frequency-inverse document frequency (TF-IDF) method is employed to analyze the classified construction report texts. The proposed CNN-based text classification model successfully classifies sentences describing construction on-site situations into six categories with an overall accuracy of 91.24%. When compared to other models, the CNN-based text classification model consistently performs with high accuracy [9].

This research proposes a feature selection method that uses the term frequency distribution measure. The Naive Bayes and SVM classifiers are employed with two benchmark datasets (WebKB and BBC). The experimental results demonstrate that the proposed feature selection method achieves higher classification accuracy. Specifically, the TFDM feature selection method reaches an accuracy of 96.7% using the NB classifier and 95.1% using the SVM classifier [10].

This paper describes the implementation of Support Vector Machines (SVM) for classifying English text and documents. Two analytical experiments were conducted to evaluate the selected classifiers using English documents. The experimental results, based on a dataset of 1,033 text documents, indicate that the Rocchio classifier performs best when the feature set is small, while SVM outperforms other classifiers. The analysis shows that the classification rate exceeds 90% when using more than 4,000 features [11].

This study employed the Doc2vec word embedding method for text classification on two datasets: the Turkish Text Classification 3600 (TTC-3600) consisting of Turkish news texts and the BBC-News dataset consisting of English news texts. Deep learning-based CNN and traditional machine learning classification methods, including Gauss Naive Bayes (GNB), Random Forest (RF), Naive Bayes (NB), and Support Vector Machine (SVM), were used as classification methods. The proposed model achieved the highest accuracy of 94.17% on the Turkish dataset and 96.41% on the English dataset using CNN classification [12].

This paper explores the potential and efficacy of Machine Learning models for text classification. Traditional machine learning methods, such as Support Vector Machines, Naïve Bayes, and Random Forests, are examined. The paper emphasizes the importance of other steps in the text classification process, including preprocessing, lemmatization, and n-gram usage, by discussing relevant datasets. The results show that the Support Vector Machine model achieved the highest accuracy rate at 95.13%, while the Random Forest algorithm using 3-grams showed the lowest accuracy [13].

This study utilizes text mining techniques and combines two supervised machine learning algorithms, Naïve Bayes and Support Vector Machines (SVM), to create a hybrid model for text classification. The hybrid model was developed using WEKA tools and Java programming language. The results indicate that the hybrid model achieved an accuracy of 96.76%, outperforming both the Naïve Bayes and SVM models, which achieved accuracies of 61.45% and

69.21%, respectively [14].

Janani introduced a new algorithm, the Optimization Technique for Feature Selection (OTFS), for text classification and compared it with commonly used classification techniques such as probabilistic neural network, support vector machine, K-nearest neighbor, and Naïve Bayes. The results show that the proposed algorithm achieves better accuracy in optimizing features and classifying text documents based on their content [2].

This paper investigates the use of capsule networks for text classification through an empirical exploration. While previous research has demonstrated the effectiveness of capsule networks in image classification, their validity in text classification is a relatively recent area of inquiry. The study achieved an accuracy rate of 74% [15]. The paper introduces the text classification process and focuses on the CNN model used in this context, which achieved a precision score of 0.85 [16]. Despite numerous proposed approaches, automated text classification continues to be an active area of research due to the imperfect effectiveness of current classifiers. Notably, few studies have utilized deep learning techniques for text classification, particularly for non-binary or multi-category classification problems. Additionally, the LSTM algorithm is a powerful type of neural network with potential applications in text classification.

Table 1 survy of literature review

Study	objective	Techniques	Accuracy
Hongxia Lu, 2022	The purpose of this study is to evaluate the performance of different deep learning algorithms in text classification tasks on medical notes related to various diseases. The dataset used in the study includes 1,237 unique discharge notes, and the task is a binary-class problem involving categorization into two classes.	CNN, LSTM, RNN, GRU, Bi-LSTM	Convolutional Neural Network 0. 93.
Thirumoorthy,2022	This study employs a feature selection method that uses the measure of term frequency distribution.	SVM., NB classifier	NB 96.7%, SVM 95.1%
Sudha, D. N. 2021.	The objective of this study is to analyze the impact of n-gram feature selection on news article text classification using semi-supervised learning methods.	support Vector Machine Naïve Bayes	SVM 93.64

Han, X,2021	The authors suggest using pre-trained language models and logic rules to create prompts with sub-prompts for multi-class text classification. This method is called prompt tuning with rules.	Prompt Tuning with Rule pre-trained language models PLMs	91.9 F1 scores
Tian, D., 2021	This study introduces an integrated approach that utilizes natural language processing technology (NLP).	CNN, LR, NB, SVM, RF, XG	CNN 91.24%, LR 89%., NB 88.32, SVM 86.67, RF 84.74, XGB 80.48
Luo, 2021	The authors used the Support Vector Machines (SVM) model for classifying English text and documents in this paper.	SVM	SVM 90%
Dogru,2021	The study employed the Doc2vec word embedding method for text classification on the Turkish Text Classification dataset.	SVM, GNB, Random Forest (RF), Naive Bayes (NB), CNN	CNN 96.41%, GNB 91,48%, RF 89.72 %, NB 91,44 %, SVM 95,05%
(Tzimourtas,2021)	This paper investigates the potential and effectiveness of Machine Learning models in text classification.	Support Vector Machine, Random Forest, Naive Bayes	SVM 95.13, RF 0.90, NB 0.94
(Asogwa,2021)	This study combines two supervised machine learning algorithms, Naïve Bayes and Support Vector Machines (SVM), with text mining techniques to create a hybrid model for text classification.	hybrid model (SVM, NB)	hybrid model gave 96.76% accuracy as against the 61.45% and 69.21% of the Naïve Bayes and SVM models respectively
(Janani, 2021)	This research introduces a novel algorithm, the Optimization Technique for Feature Selection (OTFS) algorithm, for text classification.	MLearn-ATC	0.93
(Kim,2020).	The paper conducts an empirical investigation of the utilization of capsule networks for text classification.	capsule networks	74%

(Cai, 2018).	The paper provides an introduction to text classification and specifically focuses on the convolutional neural network.	CNN	0. 85

3 Methodology & material

The main objective of this research is to classify relevant documents based on their content while minimizing time complexity. The study comprises four primary processes, namely data collection, data pre-processing, feature extraction, and the application of NLP algorithms. Figure 1 illustrates the proposed framework for document classification. In the first step, data collection was conducted from an open-source library such as Kaggle. We selected datasets on biomedical text document classification gathered from Kaggle. After data collection, the next step is preprocessing, which is detailed in the methodology section. Once preprocessing is completed, we performed feature extraction for our experiments. Our chosen features include word frequency, question marks, full stops, initial words, and final words of the documents. We utilized two classification methods, Logistic Regression (LR), LSTM, and CNN respectively, for English text mining. Each stage is described further below.

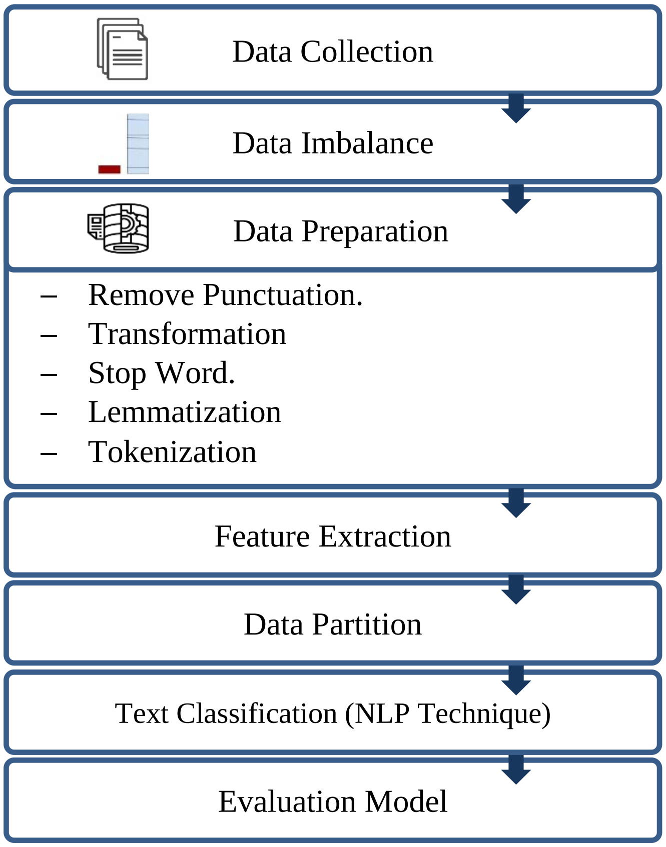
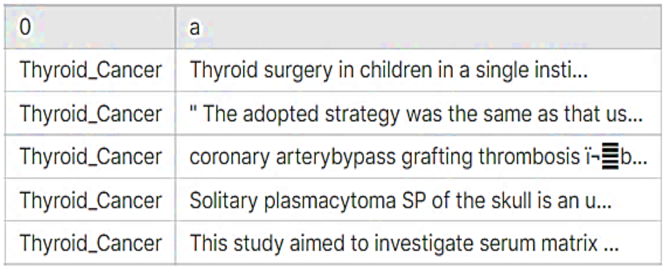


Figure 1 Overview of text classification model

1) Data Collection

For this study, we obtained a dataset on Biomedical text documents, abstracts, and full papers from the Kaggle website. We used the subset of this dataset that included long research papers with six or more pages. The dataset consisted of cancer-related documents, which were divided into three categories: Thyroid Cancer, Colon Cancer, and Lung Cancer. The total number of publications in the dataset was 7,569, and it contained three class labels. The number of samples in each category was as follows: Colon Cancer (2,579), Lung Cancer (2,180), and Thyroid Cancer (2,810).

Table 2 Sample of dataset



0	a
Thyroid_Cancer	Thyroid surgery in children in a single insti...
Thyroid_Cancer	" The adopted strategy was the same as that us...
Thyroid_Cancer	coronary arterybypass grafting thrombosis ï¬≣b...
Thyroid_Cancer	Solitary plasmacytoma SP of the skull is an u...
Thyroid_Cancer	This study aimed to investigate serum matrix ...

2) Data imbalance

When learning from class label-imbalanced data, the prediction model's accuracy may be misleading [17]. Therefore, it is important to assess whether the dataset's class labels are balanced. In our dataset, as shown in Figure 3, the number of samples for the "Colon" category was 2,579, the number of samples for the "Lung" category was 2,180, and the number of samples for the "Thyroid" category was 2,810. Based on these numbers, we can conclude that all categories are balanced.

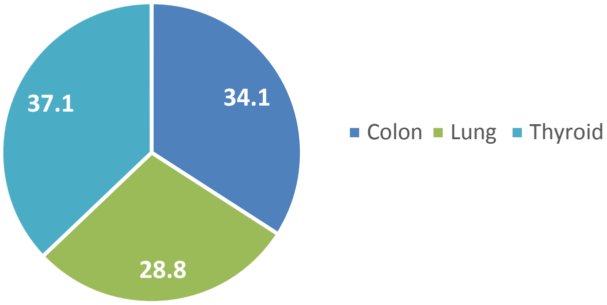


Figure 2 Visualization of Dataset Imbalance

3) Data preprocessing

Data preprocessing involves preparing and cleaning text to facilitate classification. Text and document datasets often contain unnecessary words such as stop words, misspellings, and slang. Noise and irrelevant features can negatively impact system performance, particularly in statistical and probabilistic learning algorithms [18] [12]. In this section, we employed various techniques and methods for text cleaning and preprocessing. Basic preprocessing steps were performed on the datasets before they were vectorized. These steps included the following [19].

3.1 Punctuation Removal

Punctuations such as {"?", "!", ".", ";"} hold no actual importance when it comes to the analysis of the data. So, the better practice of data analysis involves the removal of punctuation beforehand.

3.2 Text Lowercasing

The text document is transformed into the lowercase so that the uppercase and lowercase words with same meaning are not treated differently.

3.3 Stop-words Removal.

The process of categorizing texts and documents involves several words that lack significant meaning for classification algorithms, such as "a", "about", "above", "across", "after", "afterwards", "again", and so on. To handle these words effectively, they are often eliminated from the texts and documents using a common approach.

3.4 Text Lemmatization

Lemmatization refers to the process of identifying the root word, also known as the stem, from various types of words. For instance, words like 'Friendly' and 'Friends' can be reduced to a common word 'Friend' by using a suffix-stripping algorithm. This technique is widely used in text classification systems for intelligent information retrieval purposes.

3.5 Text Tokenization

Tokenization is a technique used in pre-processing of text, wherein a continuous stream of text is divided into smaller units known as tokens, such as words, phrases, symbols, etc. The primary objective of this step is to analyze the words present in a sentence. For tasks like text classification and text mining, a parser that processes the tokenization of the documents is necessary: sentence [20] [21] [2].

After sleeping for four hours, he decided to sleep for another four. In this case, the tokens are as follows:

{“After” “sleeping” “for” “four” “hours” “he” “decided” “to” “sleep” “for” “another” “four”}.

3.6 Feature Extraction

Text classification accuracy is significantly impacted by the process of text feature extraction. Feature extraction involves treating text as a dot in an N-dimensional space based on a vector space model [22].

Text Feature Extraction refers to the process of converting text into a set of features represented by a vector in real number form, which is used as input for classification purposes. In this study, the process of feature extraction from text utilized the TF-IDF model [23]. The application of Term Frequency Inverse Document Frequency (TF-IDF) weighing scheme is one of the approaches to extract features [24]. The TF-IDF weighting scheme assigns importance to a keyword based on its frequency in a document and relevance across the corpus. This process involves weighing the keyword in any context, checking how often it appears in the document and how relevant it is throughout the entire corpus. [25] [26]. The overall approach of this feature extraction method works as follows. Given a document collection D, a word w and an individual document d∈ D, we calculate:

$$w_{i,j} = tf_{i,j} \times \log(N/df_i) \quad (1)$$

$tfij$=number of occurrences of i in j, dfi=number of documents containing i, N=total number of documents.

4) Data Partition

In order to train the classification model, the dataset is divided into two non-overlapping parts for training and testing purposes. The hold-out part represents 25% of the dataset and is used for testing, while the other 75% is used for training. This portion is referred to as "hold-out" since it is held out for testing, while the remaining data is used to develop the model [27].

Table 3 show data splitting using hold-out validation method.

Category	**Input**	**Target**	**%**	**Partition**
Part 1	6056	(6056, 3)	75%	Data Training
Part 2	1514	(1514) **Colon** =514, **Lung**=433, **Thyroid** =567	25%	Hold-out/ Test set

5) Text Classification Algorithms

At this stage, we applied some natural language algorithms in the classification process, namely Logistic Regression, long short-term memory (LSTM) and CNN. As indicated in figure, the model was subsequently trained using the training data selected in the previous phase (data splitting)

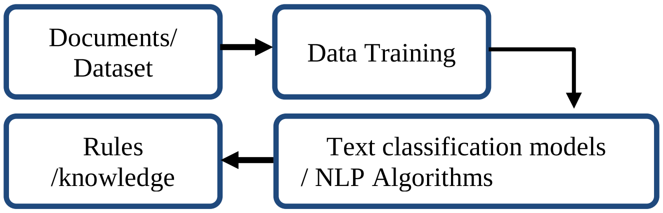


Figure 3 traning model process

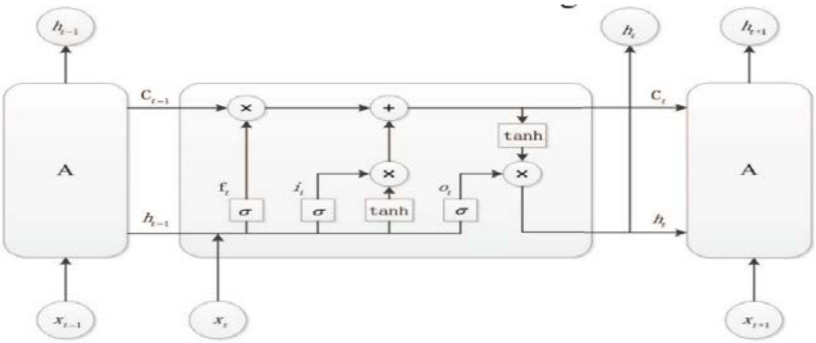
a) Logistic regression

Logistic regression is a supervised classification algorithm that has gained significant importance in recent times. This algorithm works by taking input and multiplying the input value with weight value. It is a type of classifier that learns which features from the input are most useful for distinguishing between different possible classes [28].

b) Convolutional Neural Network

The Convolutional Neural Network (CNN) provides the best feature extraction techniques through various layers, including embedding, convolutional, pooling, and fully connected layers. The embedding layer transforms preprocessed text by converting each word into an embedding vector with a specific dimension. The convolutional layer reduces the dimensionality of features or phrases, while the pooling layer merges similar features together. The flatten layer then transforms the 2D features into a vector form suitable for the fully connected layer. Dropout and activation functions are used in one or more neural layers, called the fully connected or dense layer, to train the model. The output layer, which has nodes equal to the number of classes in the dataset, uses the SoftMax activation function to predict models. Another technique used is Long Short-Term Memory (LSTM). [29].

LSTM has the characteristics of long-distance context dependent learning and can store context history information; thus, it is used as timing layer of the model in this paper. The module structure



of LSTM is shown in Fig. 4. [30] [31].

Figure 4 Structure of LSTM

6) Evaluation Model & Results

For multiclass classification, performance measures like Confusion Matrix, Accuracy, Precision, Recall, and F1-Score are defined based on four features: true positive (TPi), true negative (TNi), false positive (FPi), and false negative (FNi) of class Ci. If there are m classes in the dataset, i ranges from 1 to m. There are three ways to calculate precision, recall, and F1-score over the entire test data: macro-averaged, micro-averaged, and weighted-averaged. In this experiment, accuracy, as well as precision, recall, and F1-score based on weighted-average, are used to evaluate the classifier's performance. These performance measures are defined as shown in [11] [28][29] [32].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

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

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$F1.Score = 2 * \frac{Precision * recall}{precision + recall} \quad (5)$$

Figure 5 confusion matrix of LSTM

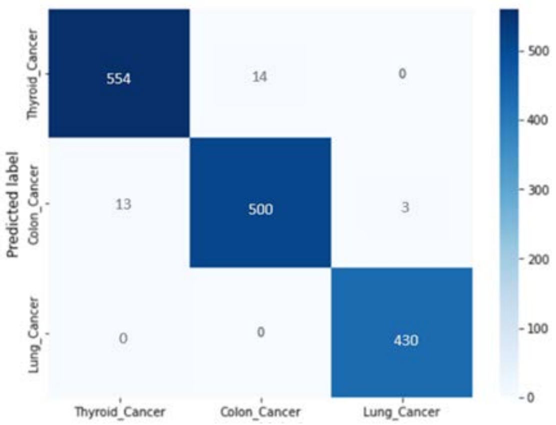


Figure 6 Confusion Matrix of LR

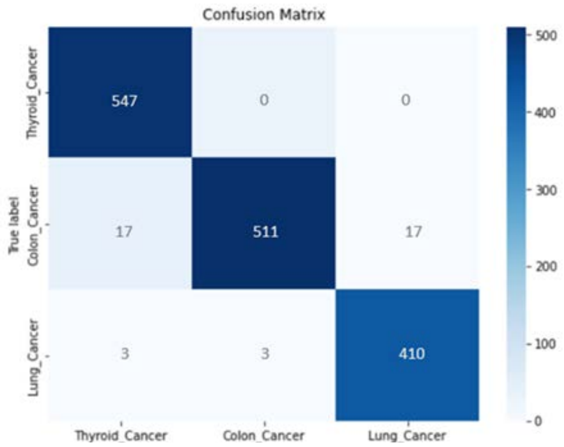


Figure 7 Confusion Matrix of CNN

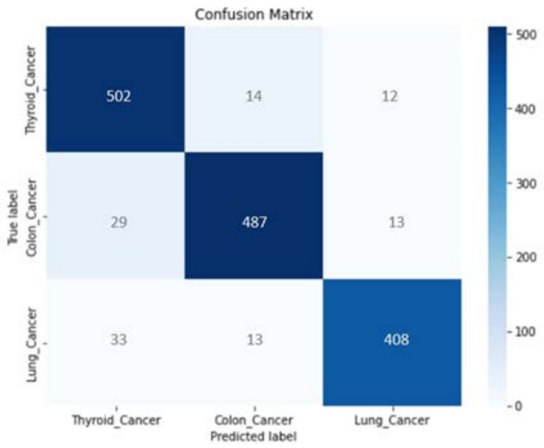


Table 3 Accuracy of models

	Accuracy	*Precision*	*Recall*
LSTM	0.980	0.984	0.982
CNN	0.969	0.965	0.964
LR	0.924	0.925	0.923

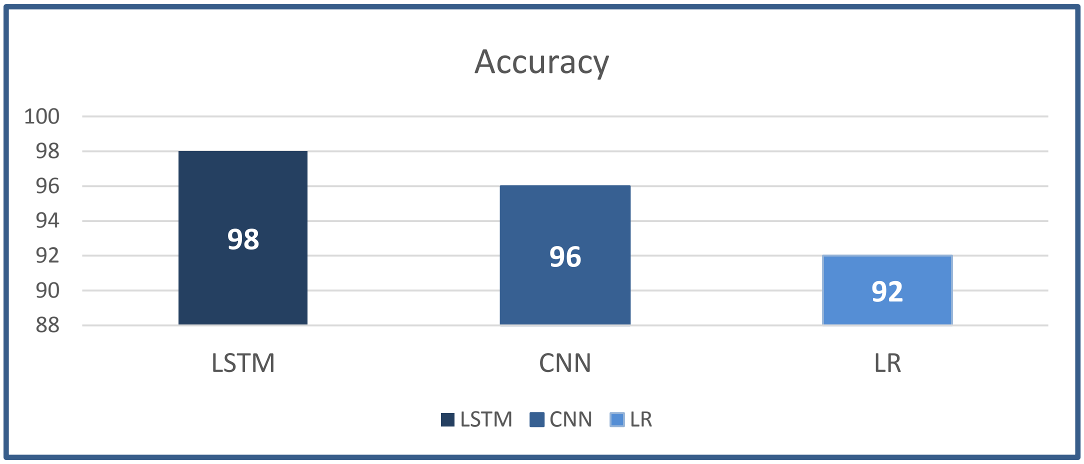


Figure 8 visualization of techniques results

From the results in Table 4 and Figure 8, we have calculated the accuracy precision, and recall value based on our simulation. The equations are provided in Equations. (2), (3), and (4) respectively. From Table 4 and Figure it's obvious that LSTM techniques provide more efficiency as compared to CNN and Logistic Regression, each technique achieved an accuracy of 0.9940.973, 0.954 respectively. When comparing the results of the proposed model and previous studies that used the same techniques, it became clear that the accuracy of the proposed model is higher than the accuracy of the models of previous studies, as shown in Table 5.

Table 4 comparsion between previous studies and proposed models

Studies	**Techniques**		
	LSTM	CNN	LR
Hongxia Lu, 2022	75%	93%	-
Tian, D., 2021	-	91.24%,	89%
Dogru,2021	-	96.41%,	-
(Cai, 2018).	-	85%	-
Proposed model	98%	96.9%	92.3%

Conclusion

Discharge medical notes written by physicians that contain important information about the health condition of patients. Many deep learning algorithms have been successfully applied to extract important information from unstructured medical notes data that can entail subsequent actionable results in the medical domain. This study aims to explore the model performance of various deep learning algorithms in text classification tasks on medical notes, in this study, we employed 3 artificial intelligence models, CNN (Convolutional Neural Network), LSTM (Long short-term memory), and Logistic regression. The analysis of these 3 categories of classification problems showed that the LSTM model performs the best.

REFERENCES

[1] Sarkar, A., & Datta, D. (2017). A Frequency Based Approach to Multi-Class Text Classification. *International Journal of Information Technology and Computer Science (IJITCS)*, *9*(5), 15-22.

[2] Janani, R., & Vijayarani, S. (2021). Automatic text classification using machine learning and optimization algorithms. *Soft Computing*, *25*(2), 1129-1145.

[3] Lee, J., Song, H. J., Yoon, E., Park, S. B., Park, S. H., Seo, J. W., ... & Choi, J. (2018). Automated extraction of Biomarker information from pathology reports. *BMC medical informatics and decision making*, *18*(1), 1-11.

[4] Santos, T., Tariq, A., Gichoya, J. W., Trivedi, H., & Banerjee, I. (2022). Automatic Classification of Cancer Pathology Reports: A Systematic Review. *Journal of Pathology Informatics*, *13*, 100003.

[5] Zhang, X., Zhang, Y., Zhang, Q., Ren, Y., Qiu, T., Ma, J., & Sun, Q. (2019). Extracting comprehensive clinical information for breast cancer using deep learning methods. *International Journal of Medical Informatics*, *132*, 103985.

[6] Lu, H., Ehwerhemuepha, L., & Rakovski, C. (2022). A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. BMC Medical Research Methodology, 22(1), 181.

[7] Sudha, D. N. (2021). Semi Supervised Multi Text Classifications for Telugu Documents. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(12), 644-648.

[8] Han, X., Zhao, W., Ding, N., Liu, Z., & Sun, M. (2021). Ptr: Prompt tuning with rules for text classification. *arXiv preprint arXiv:2105.11259*.

[9] Tian, D., Li, M., Shi, J., Shen, Y., & Han, S. (2021). On-site text classification and knowledge mining for large-scale projects construction by integrated intelligent approach. *Advanced Engineering Informatics*, *49*, 101355.

[10] Thirumoorthy, K., & Muneeswaran, K. (2022). Feature selection for text classification using machine learning approaches. *National Academy Science Letters*, *45*(1), 51-56.

[11] Luo, X. (2021). Efficient english text classification using selected machine learning techniques. *Alexandria Engineering Journal*, *60*(3), 3401-3409.

[12] Dogru, H. B., Tilki, S., Jamil, A., & Hameed, A. A. (2021, April). Deep learning-based classification of news texts using doc2vec model. In *2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA)* (pp. 91-96). IEEE.

[13] Tzimourtas, A., Bakalakos, S., Tselenti, P., & Voulodimos, A. (2021, November). An exploration on text classification using machine learning techniques. In *25th Pan-Hellenic Conference on Informatics* (pp. 247-249).

[14] Asogwa, D. C., Anigbogu, S. O., Onyenwe, I. E., & Sani, F. A. (2021). Text classification using hybrid machine learning algorithms on big data. *arXiv preprint arXiv:2103.16624*.

[15] Kim, J., Jang, S., Park, E., & Choi, S. (2020). Text classification using capsules. *Neurocomputing*, *376*, 214-221.

[16] Cai, J., Li, J., Li, W., & Wang, J. (2018, December). Deeplearning model used in text classification. In *2018 15th international computer conference on wavelet active media technology and information processing (ICCWAMTIP)* (pp. 123-126). IEEE.

[17] Garcia, S., Luengo, J., Sáez, J. A., Lopez, V., & Herrera, F. (2012). A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE transactions on Knowledge and Data Engineering*, *25*(4), 734-750.

[18] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, *10*(4), 150.

[19] Singh, G., Kumar, B., Gaur, L., & Tyagi, A. (2019, April). Comparison between multinomial and Bernoulli naïve Bayes for text classification. In *2019 International Conference on Automation, Computational and Technology Management (ICACTM)* (pp. 593-596). IEEE.

[20] Gupta, G.; Malhotra, S. Text Document Tokenization for Word Frequency Count using Rapid Miner (Taking Resume as an Example). Int. J. Comput. Appl. 2015, 975, 8887.

[21] Verma, T.; Renu, R.; Gaur, D. Tokenization and filtering process in RapidMiner. Int. J. Appl. Inf. Syst. 2014, 7, 16–18. [CrossRef]

[22] H. Liang, X. Sun, Y. Sun and Y. Gao, "Text feature extraction based on deep learning: a review," EURASIP journal on wireless communications and networking, vol. 2017, no 1, 211, 2017.

[23] Indra, S. T., Wikarsa, L., & Turang, R. (2016, October). Using logistic regression method to classify tweets into the selected topics. In 2016 international conference on advanced computer science and information systems (icacsis) (pp. 385-390). IEEE.

[24] Dzisevič, R., & Šešok, D. (2019, April). Text classification using different feature extraction approaches. In 2019 Open Conference of Electrical, Electronic and Information Sciences (eStream) (pp. 1-4). IEEE.

[25] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF," Journal of documentation, vol. 60, no. 5, pp. 503-520, 2004

[26] Liu, Q., Wang, J., Zhang, D., Yang, Y., & Wang, N. (2018, December). Text features extraction based on TF-IDF associating semantic. In 2018 IEEE 4th International Conference on Computer and Communications (ICCC) (pp. 2338-2343). IEEE.

[27] Yadav, Sanjay, and Sanyam Shukla. "Analysis of k-fold cross-validation over holdout validation on colossal datasets for quality classification." 2016 IEEE 6th International conference on advanced computing (IACC). IEEE, 2016.

[28] Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augmented Human Research, 5(1), 1-16.

[29] Behera, B., Kumaravelan, G., & Kumar, P. (2019, December). Performance evaluation of deep learning algorithms in biomedical document classification. In 2019 11th international conference on advanced computing (ICoAC) (pp. 220-224). IEEE

[30] Zhou, C., Sun, C., Liu, Z., & Lau, F. (2015). A C-LSTM neural network for text classification. arXiv preprint arXiv:1511.08630.

[31] Bai, X. (2018, September). Text classification based on LSTM and attention. In 2018 Thirteenth International Conference on Digital Information Management (ICDIM) (pp. 29-32). IEEE

[32] Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. WSEAS transactions on computers, 4(8), 966-974.