Available online at www.medicinescience.org

ORIGINAL ARTICLE

Medicine Science International Medical Journal

Medicine Science 2024;13(1):171-4

Comparative analysis of machine learning algorithms for biomedical text document classification: A case study on cancer-related publications

Ekrem Kucuk[1,2], Ipek Balikci Cicek[1], Zeynep Kucukakcali[1], Cihan Yetis[3]

[1]İnönü University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Türkiye
[2]Carbon Health, Senior Software Engineer, Ankara, Türkiye
[3]Kilis Prof. Dr. Alaeddin Yavaşca State Hospital, Department of Cardiovascular Surgery, Kilis, Türkiye

Received 11 October 2023; Accepted 27 December 2023
Available online 16.02.2024 with doi: 10.5455/medscience.2023.10.209

Abstract

Biomedical text document classification is an essential task within Natural Language Processing (NLP), with applications ranging from sentiment analysis to authorship identification. Despite advancements in traditional machine-learning algorithms like Support Vector Machines (SVM) and Logistic Regression, challenges such as data sparsity and high dimensionality persist. Recent years have seen a surge in the use of deep learning models to mitigate these issues. This study aims to conduct a comparative analysis of various machine-learning algorithms for classifying biomedical text documents. The study employs the "Medical Text Dataset - Cancer Doc Classification" from Kaggle, comprising 7570 biomedical text documents labeled into three types of cancer (colon, lung, and thyroid). A preprocessing pipeline involving tokenization, stop-word removal, and Term Frequency-Inverse Document Frequency (TF-IDF) vectorization is applied. Algorithms including Logistic Regression, SVM, and Multinomial Naive Bayes are evaluated through 5-fold cross-validation. Performance metrics like accuracy, precision, recall, F1 score, and area under the ROC curve (AUC ROC) are employed. Logistic Regression outperforms the other algorithms with an accuracy of 78.3% and an AUC ROC of 88.59%. SVM and Multinomial Naive Bayes follow with lower performance metrics. Hyperparameter tuning further enhances the performance of the algorithms, particularly Logistic Regression. The study makes a significant contribution to the field of biomedical text classification by systematically comparing machine-learning algorithms. Logistic Regression emerges as the most effective, emphasizing the importance of algorithm selection and hyperparameter tuning in machine learning applications within this domain.

Keywords: Artificial intelligence, biomedical text classification, machine learning, natural language processing

Introduction

Biomedical text document classification, also known as biomedical text document categorization when considered in the English linguistic context, constitutes a pivotal task within the expansive domain of Natural Language Processing (NLP). This computational endeavor facilitates the categorization of text documents into predefined classes or categories, contingent upon the semantic and syntactic attributes of their content. The significance of biomedical text document classification is underscored by its wide-ranging applications, which span from web content categorization and unstructured text classification to sentiment analysis, spam e-mail filtration, and even authorship identification [1].

The methodology for biomedical text document classification generally involves a series of sequential steps. Initially, the process commences with tokenization, which entails the disintegration of the text into smaller, manageable units, commonly referred to as tokens. These tokens could be words, phrases, or even punctuation marks. Subsequently, the algorithm proceeds to eliminate superfluous keywords or stop words are typically conjunctions, prepositions, pronouns, and articles that do not substantially contribute to the semantic richness of the text. The third step involves the computation of term frequency and the assignment of weight to each term, thereby quantifying its significance within the document or the class it represents. Upon completion of these preprocessing steps, the text documents are

CITATION
Kucuk E, Balikci Cicek I, Kucukakcali Z, Yetis C. Comparative analysis of machine learning algorithms for biomedical text document classification: A case study on cancer-related publications. Med Science. 2024;13(1):171-4.

Corresponding Author: Ipek Balikci Cicek, İnönü University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Türkiye
Email: ipek.balikci@inonu.edu.tr

DOI: 10.5455/medscience.2023.10.209

transmuted into vector space models, which are then fed into a classifier for the final categorization [2].

A plethora of machine-learning algorithms can serve as classifiers in this context. These range from traditional algorithms like Support Vector Machines (SVM) and Logistic Regression Analysis to ensemble methods like Gradient Boosting Trees (GBT) and eXtreme Gradient Boosting (XGBoost). The efficacy of these algorithms is commonly assessed through a variety of performance metrics, including but not limited to, accuracy, precision, recall, and the F-measure [3].

However, the task is not devoid of challenges. Some of the most pressing issues include data sparsity, high dimensionality, the presence of noisy data, imbalanced data distribution, and the utilization of domain-specific terminologies. To surmount these obstacles, deep learning models have emerged as a promising avenue in recent years. These models are capable of autonomously learning both the semantic and syntactic features embedded within the text documents, thereby enhancing classification performance [4].

The primary objective of this research paper is to conduct a comparative analysis of the performance metrics of various machine-learning algorithms in the context of biomedical text document classification. To achieve this aim, we have employed a diverse set of machine-learning algorithms, training and testing them on a publicly available biomedical text classification dataset. By addressing these facets, this paper endeavors to contribute to the existing body of knowledge by identifying the most efficacious machine-learning model for biomedical text document classification.

Material and Methods

Since the data set used in this study is open access, ethics committee approval is not required.

Dataset utilization

In the context of the current investigation, a publicly available dataset, denominated as "Medical Text Dataset-Cancer Doc Classification Cancer Text Documents Classification," was employed as the foundational data source. The public dataset comprising 7570 biomedical text documents is accessible via the Kaggle platform and serves as an exhaustive repository of scholarly text documents that are specifically oriented toward topics related to cancer research. Each document within this curated dataset is scrupulously labeled to correspond with one of three distinct types of cancer: colon, lung, and thyroid. The meticulous labeling process enhances the dataset's utility for machine learning applications. The overarching aim of leveraging this particular dataset is to facilitate the systematic development, training, and rigorous evaluation of text classification algorithms. The ultimate objective is to engender models that are proficient in categorizing the documents with a high degree of accuracy, specifically based on the type of cancer being discussed in the text [5].

Data preprocessing

Prior to the deployment of machine learning algorithms for textual analysis, a comprehensive series of preprocessing steps were meticulously executed on the corpus of text documents. This preprocessing pipeline was designed to optimize the text data for subsequent analytical procedures. The initial phase involved the standardization of text by converting all alphabetic characters to lowercase, thereby ensuring uniformity across the dataset. Subsequently, all punctuation marks, numerical figures, and extraneous spaces or characters were systematically eliminated to reduce noise and enhance the quality of the data. In addition to these measures, a list of English stop words—commonly occurring words that contribute little to semantic meaning—were identified and excluded from the dataset. To further refine the text, stemming techniques were employed to truncate words to their root forms, thereby consolidating variations of the same term. Upon the completion of these preprocessing activities, Term Frequency-Inverse Document Frequency (TF-IDF) vectors were generated for each individual document. The TF-IDF algorithm is a widely acknowledged and robust method for quantifying the relative significance of individual words within a corpus, thereby facilitating more accurate and insightful machine learning analyses [6,7].

Model selection and evaluation

In the pursuit of effective text classification, an array of machine learning algorithms was rigorously investigated for their suitability in this specialized domain. The algorithms under consideration encompassed Logistic Regression, SVM, and Multinomial Naive Bayes. Naive Bayes is an efficient and expeditious method that demonstrates proficiency in several text categorization predicaments. The underlying premise is that the words in a text are unrelated to one another, and that the likelihood of a text belonging to a certain category is directly proportional to the multiplication of the probabilities of each word in that category. Naive Bayes is capable of effectively managing extensive and sparsely populated data sets, as well as effectively handling many classes. However, its performance may be suboptimal in cases when the words are not independent or when there are significant correlations between characteristics and classes. In order to employ Naive Bayes for text classification, it is necessary to initially transform the text into a vector representation consisting of word counts or frequencies. Subsequently, the Bayes theorem is applied to compute the probability of each class. Logistic Regression is another popular and versatile algorithm that can be used for text classification. It is a linear model that predicts the probability of a text belonging to a class by using a logistic function. Logistic Regression can handle both binary and multiclass problems, and can also incorporate regularization techniques to prevent overfitting. Logistic Regression can capture the linear relationships between the words and the classes, but it may not be able to capture the complex and nonlinear patterns in the text. To use Logistic

DOI: 10.5455/medscience.2023.10.209

Regression for text classification, you need to first convert your text into a vector of word counts or frequencies, or use a more advanced technique like TF-IDF, and then fit the logistic function to the data. SVMs are robust and versatile algorithms that are applicable for text categorization. SVMs are founded on the concept of identifying the most favorable hyperplane that effectively separates data points belonging to distinct classes, while maximizing the distance between them. SVMs are capable of addressing both linear and nonlinear issues. Additionally, SVMs have the ability to employ various kernels to convert the data into higher-dimensional spaces. SVMs may attain elevated levels of accuracy and generalization. However, they can also incur significant computing costs and exhibit sensitivity to the selection of parameters and kernels. In order to employ SVMs for text classification, it is necessary to initially transform the text into a vector representation based on either word counts or frequencies. Alternatively, a more sophisticated approach such as TF-IDF can be utilized. Subsequently, the SVM model must be trained using the suitable kernel and parameters.

To ensure the robustness of the models, a 5-fold cross-validation methodology was employed during the training phase. Additionally, hyper-parameters were meticulously optimized through the utilization of a Grid Search algorithm. To rigorously assess the performance of these models, a comprehensive set of evaluation metrics was employed. These metrics included accuracy, precision, recall, F1 score, and the area under the Receiver Operating Characteristic curve (AUC ROC) [8].

Results

This research employed various machine learning algorithms to classify scientific publications into distinct categories of diseases—Colon Cancer, Lung Cancer, and Thyroid Cancer. The algorithms utilized were Logistic Regression, Support Vector Machine, and Multinomial Naive Bayes. The performance of the models was assessed using 5-fold cross-validation on the dataset. The performance metrics used for evaluation included Accuracy, Precision, Recall, F1 score, and AUC ROC, as pointed out in Table 1.

Table 1. The performance metrics results of the constructed models

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC ROC (%)
Logistic regression	78.3	78.83	78.3	78.09	88.59
Support vector machine	75.11	75.81	75.11	75.01	85.21
Multinomial naive bayes	65.06	64.93	65.06	64.49	81.48

The process of hyperparameter tuning led to substantial improvements in the performance metrics. The Logistic Regression model, when tuned, achieved an accuracy of approximately 78.30% and an AUC ROC of 88.59%, which were the highest among all models. Logistic Regression demonstrated the highest performance across most of the metrics. It achieved an accuracy of 78.3%, a precision of 78.83%, a recall of 78.3%, an F1 score of 78.09%, and an AUC ROC of 88.59%. These results suggest that Logistic Regression is highly effective in both classifying the biomedical text documents accurately and distinguishing between the different classes. The AUC ROC score provided deeper insights into each model's classification performance. Logistic Regression outperformed the other models with an AUC ROC of 88.59%.

SVM followed closely, with an accuracy of 75.11%, a precision of 75.81%, a recall of 75.11%, an F1 score of 75.01%, and an ROC AUC of 85.21%. While SVM performed well, it was slightly less effective than Logistic Regression in all metrics except for precision, where it was marginally lower.

Multinomial Naive Bayes lagged behind the other two algorithms, with an accuracy of 65.06%, a precision of 64.93%, a recall of 65.06%, an F1 score of 64.49%, and an AUC ROC of 81.48%. The algorithm was notably less effective in classifying biomedical text documents, as evidenced by its lower scores across all metrics.

Discussion

The central objective of the present research endeavor is to conduct a meticulous assessment and comparative analysis of the performance metrics associated with a diverse array of machine-learning algorithms, specifically within the specialized domain of biomedical text classification. To achieve this ambitious goal, an extensive selection of machine-learning models has been employed, each distinguished by its unique computational architecture and underlying theoretical principles. These selected models were subjected to rigorous training and validation procedures, utilizing a publicly available dataset that embodies the complex challenges inherent to the realm of biomedical textual data. This dataset serves as a representative microcosm for the intricacies and nuances that typify biomedical literature. Through the implementation of this comprehensive research methodology, the study aspires to make a substantive contribution to the existing body of academic literature. It aims to elucidate the relative strengths and weaknesses of each computational approach under investigation. By doing so, the research seeks to identify the most efficacious machine-learning model for tasks related to biomedical text classification. This scholarly undertaking holds significant potential for advancing the current state of the art, by providing empirically substantiated insights into the algorithmic efficacy and practical applicability of machine-learning techniques in this highly specialized field.

DOI: 10.5455/medscience.2023.10.209

The Logistic Regression model's superior performance aligns with previous research that has also identified its efficacy in text classification tasks [9]. The model achieved an accuracy rate of 78.30% and an AUC ROC score of 88.59%, indicating its robustness in classifying biomedical documents. In a similar vein, a study by Wang and Manning [10] demonstrated that Logistic Regression models are well-suited for high-dimensional text data, corroborating the findings of the current research.

The Support Vector Machine model followed closely, albeit with slightly lower performance metrics. This result is consistent with studies that have found SVM to be effective, but often less so than Logistic Regression in text classification tasks [11]. However, it is worth noting that SVM had a marginally lower precision compared to Logistic Regression, which suggests that SVM may be more conservative in making positive identifications. A study by Forman [12] posits that SVM can sometimes be overly cautious, thereby affecting its precision rate.

The Multinomial Naive Bayes model lagged behind, corroborating existing literature that highlights its limitations in handling high-dimensional and imbalanced datasets [13]. The model achieved an accuracy rate of 65.06% and an AUC ROC score of 81.48%, both of which are substantially lower than those achieved by Logistic Regression and SVM. It is worth noting that the Naive Bayes classifiers are often considered a baseline method in text classification [9], and the current study reinforces this perspective.

The process of hyperparameter tuning led to substantial improvements in performance metrics for all algorithms, especially for Logistic Regression. This is in line with research by Bergstra and Bengio [14], which emphasizes the importance of hyperparameter tuning in machine learning models.

While the current study contributes valuable insights, there are several limitations to consider. First, due to computational constraints, the study relied on 5-fold cross-validation on a subset of the dataset. Previous studies have emphasized the benefits of using larger datasets and more complex cross-validation techniques [15]. Secondly, the study focused only on three categories of diseases, which limits the generalizability of the findings. Future research could benefit from incorporating a wider range of biomedical text documents to assess the scalability and adaptability of the algorithms.

Conclusion

The present study makes a meaningful contribution to the literature on machine learning applications in biomedical text classification. The findings indicate that Logistic Regression is the most effective algorithm among those tested, followed by SVM and Multinomial Naive Bayes. These results underscore the need for careful algorithm selection and hyperparameter tuning in machine learning projects, particularly those focused on biomedical text classification.

Conflict of Interests
The authors declare that there is no conflict of interest in the study.

Financial Disclosure
The authors declare that they have received no financial support for the study.

Ethical Approval
Given that the data set utilized in this study is openly accessible, there is no need for permission from an ethical committee.

References

1. Aggarwal CC, Zhai C. A survey of text classification algorithms. In: Mining text data. 2nd edition. Springer, Boston, 2012;163-222.
2. Manning CD, Surdeanu M, Bauer J, et al. The Stanford CoreNLP natural language processing toolkit. Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System DemonstrationsAt: Baltimore, Maryland, 2014, 55-60.
3. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, 13-17 August 2016, 785-794.
4. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436-44.
5. Medical text dataset -cancer doc classification. https://www.kaggle.com/datasets/falgunipatel19/biomedical-text-publication-classification access date 10.11.2023
6. Schütze H, Manning CD, Raghavan P. Introduction to information retrieval. In: Text classification and Naive Bayes. 3rd edition.Cambridge University Press, Cambridge, 2008;253-288.
7. Falasari A, Muslim MA. Optimize naïve bayes classifier using chi square and term frequency inverse document frequency for amazon review sentiment analysis. JOSCEX. 2022;3:31-6.
8. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manage. 2009;45:427-37.
9. McCallum A, Nigam K. A comparison of event models for naive bayes text classification. AAAI-98 workshop on learning for text categorization; 1998: Madison, WI.
10. Wang SI, Manning CD. Baselines and bigrams: Simple, good sentiment and topic classification. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics Jeju Island, Korea, July 8-14, 2012, 90-94.
11. Joachims, T. Text categorization with support vector machines: learning with many relevant features. In: Nédellec C, Rouveirol C. editors. Machine Learning: ECML-98. ECML 1998. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 1998.
12. Forman G. An extensive empirical study of feature selection metrics for text classification. J Mach Learn Res. 2003;3:1289-305.
13. Rennie JD, Shih L, Teevan J, Karger DR. Tackling the poor assumptions of naive bayes text classifiers. Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.
14. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. JMLR. 2012;13:281-305.
15. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai; 1995: Montreal, Canada. 1995, 1137-1145.