Journal of Information Systems Engineering and Management
2025, 10(25s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Predicting Sentiments of Users about Medical Treatments using Pre-trained Large Language Models

Alaa Hassan [a], Jamshid Bagherzadeh Mohasefi [a], Amir Sorayaie Azar [a]
[a] *Department of Computer Engineering, Urmia University, Urmia, Iran*

Corresponding author: it.alaa2010@coart.uobaghdad.edu.iq

ARTICLE INFO

Received: 19 Dec 2024
Revised: 10 Feb 2025
Accepted: 22 Feb 2025

ABSTRACT

Users of social media write their opinions about products, services, market, social life, health and any facet of the life in the form of texts in web-based or mobile applications. The other users use these comments to select the best services and products. In the area of the medication, the production of user generated texts has been increased, because of information explosion and technological advancements. Manual extraction of useful knowledge using the tremendous amount of textual data is impossible. Opinion mining and Sentiment Analysis (SA) is a crucial mechanism for extracting useful knowledge, including users' opinions about medical systems, to help physicians with this information. Physicians will use the extracted information to know how patients feel about the course of treatment and other health related topics. This paper investigates the application of Large Language Models (LLMs) to predict polarity of patients' opinions. This study uses a dataset that includes patient reviews regarding their opinions about medications, prescriptions, and treatment. Three scenarios are considered in this paper: scenarios of two classes (positive, negative), three classes (positive, neutral, negative), and five classes (negative, slightly negative, neutral, slightly positive, positive). BERT and DistilBERT tokenization methods are used for word embedding. For training and fine tuning in clinical domains, one traditional ML based method, One Boosting based method, and three BERT-based methods, are utilized in model development. We found the best hyper-parameters for all models using Grid-CV method. The results reveal that the fine-tuned BERT model with corresponding word embedding representation, achieved the best results, with accuracy and F1-Score of 97.71% and 98% in two classes, 97.24% and 97% in three classes, and 80.35% and 80% in five classes, respectively. Due to the high accuracy, the proposed models can be used as an auxiliary tool in clinics and medical centers.

Keywords: Machine Learning, Large Language Models, Text Analysis, Explainable Artificial Intelligence, Patients Opinion Mining, BERT, DistilBERT, LLMWare

1. INTRODUCTION

One of the easiest and straightforward data generation methods available in the web is text data. The massive amount of text data, which is unstructured, makes a potential source for knowledge discovery. Various information could be extracted from unstructured text: users' opinions, sentiments, emotions, Named Entity Recognition (NER), structured data extraction, and many other useful information. Recently, more investigation has been put on text mining task due to massive volume of unstructured text data, generated by web or mobile applications [1].

Over the past years notable increase has been happened in user comments evaluations on websites and information systems. Users review a variety of products on associated websites, such as marketing, clinical services, home appliances, movies, dining establishments, and pharmaceuticals. Users of web contribute with their comments, in web-forums, feedback forms, and review websites. Before using or buying goods or services, people read existing reviews about them, which help them make better decisions based on prior opinions. On the other hand, in order to produce better outcomes, contemporary organizations need to utilize user-generated content [2]. To extract users' opinions, much research is being done usually denoted as opinion mining or Sentiment Analysis (SA). In SA, researchers focus on determining the polarity of textual data [3]. There are different levels in extracting sentiments:

The polarity of a complete document, the polarity of a sentence in Therse are called document level, sentence level, and aspect level respectively [3]. Numerous algorithms as designed to extract sentiments most of them being classification-based tasks [4]. Movies reviews sentiment analysis, customers feedback analysis in marketplace products, hotel reviews analysis, and assessments of tourist site experiences, are just some of the common examples in this field [5, 6].

Medical texts such as users' comments on the drugs' effects, treatment quality, drugs' side-effect, patient opinions and opinions on other medical services are growing nowadays [4, 7, 8]. Before giving a medication, doctors can get automatic hint about the effects of their medication, from the tools analyzing reviews on those medications. These SA tools can give useful information to physicians and pharmacists about patients' experiences regarding the efficacy of various medications [4, 7, 8]. It is feasible to employ patient sentiment prediction in medicine to assist with future therapy because the results of different medical treatments and awareness of their usefulness have been researched.

Numerous studies have already employed machine learning and deep learning techniques to identify sentiments in non-clinical contexts [10]. However, prior clinical and medical research has a number of shortcomings [4, 10, 11–12], including the following:

- Studies rarely have considered hyper-parameter tuning in focus.
- Rarely large datasets are considered in the studies in this field.
- There are limited number of studies for identifying sentiments in medical area.
- No study has compared the various versions of the Large Language Model (LLM), Bidirectional Encoder Representations from Transformer (BERT) models, as we have done here, on the drug dataset.

In order to overcome these constraints, this work aims to provide a novel method for fully assessing the sentiments of pharmaceutical reviews. For the purpose of predicting the sentiment and opinions of patient reviews on drugs, three machine learning, three boosting-based, and three LLM-based models have been built. In summary, our main contributions to this work are as follows:

We implemented, trained, and compared three ML-based, three Boosting based and three LLM-based models.

- We considered three different approaches to study patients' sentiment classes and ratings.
- Various new preprocessing techniques have been done to prepare the data to be used in ML, Boosting, and LLM based methods.
- Grid Search has been used to identify and choose the optimal values for the ML and DL models' hyper-parameters for prediction.

This paper includes five sections. An overview of earlier research on SA is given in Section 2, followed by a description of the materials and techniques used in this study in Section 3, the results in Section 4, and a conclusion and future work plan in Section 5.

2. RELATED WORK

In an effort to better understand patients' needs, sentiments, and situations as well as the drawbacks of each treatment, numerous researchers have attempted to look at patient reviews of prescription drugs [4, 10, 11–17]. SA can be quite helpful in this area to extract relevant data for the previously indicated goal. The majority of SA researchers have used traditional and rule-based machine learning techniques [16]. SA was considered In drug reviews, using traditional machine learning techniques such as Naïve Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), and K Nearest Neighbor (KNN) [13, 15, 17, and 18]. Furthermore, novel feature extraction methods and machine learning algorithms have been proposed to improve the performance of models in SA of pharmaceutical reviews [18].

Authors in [16] suggested a sentiment extraction and recognition method using SVM and rule-based algorithms to find drug side effects in user reviews. The drug reviews for the model's training and testing were manually chosen by a medical professional. According to their findings, the SVM algorithm performed noticeably better than rule-based methods. Similar to this, [13] used LR for sentiment prediction in both cross-domain and in-domain SA,

emphasizing drug side effects, effectiveness, and overall satisfaction. They gathered information via crawling Medications and Druglib websites and obtained an accuracy of 70.06% in cross-domain SA. For predicting side effects, the accuracy ratings was 49.75%, respectively. Authors in [15] used DT, NB, and RF algorithms to apply SA to drug reviews, classifying sentiments as neutral, positive, or negative. They used fuzzy-rough feature selection for dimensionality reduction in order to overcome the difficulty of high-dimensional feature space. To measure term importance, they employed the Bag of Words (BoW) and Term Frequency, Inverse Document Frequency (TF-IDF) approaches. The model with the highest accuracy, 66.41% was RF.

Lexicon-based and supervised learning SA approaches were used in a different study [17] to examine patient emotions about pharmaceutical and medical issues. Their dataset was gathered from discussion boards online. According to the results, it was difficult to separate views about pharmaceuticals from those concerning medical professionals. Authors in [18] suggested a unique feature extraction technique using position embedding, for sentiment extraction of drug reviews. For sentiment classification, they used a variety of machine learning models and feature extraction techniques, and they showed that their suggested strategy performed better than the competitors. In a parallel study, [10] sought to use patient reviews to predict how patients will feel about their drugs. They accomplished this by using a convolutional neural network (CNN) for classification, which outperformed more conventional techniques like SVM, based on their findings.

Using both traditional ML and DL models, authors in [4] presented two unique methods for gleaning sentiments from patient medication evaluations. When they contrasted DL-based models with conventional machine learning techniques, they found that their top-performing strategy increased accuracy by 4%.

Several holes in SA for medication reviews have been found based on these studies and our own research. Online drug reviews are a vital resource for physicians, offering insightful information on patients' diseases and adverse drug reactions [19]. However, a thorough and domain-specific sentiment lexicon for clinical pharmaceutical reviews has not yet been developed by previous research [4, 10, 12-18, 20]. Despite being a valuable resource for SA, the dataset utilized in [13] has not yet been thoroughly examined. Furthermore, the use of different pre-trained Transformers for SA in pharmaceutical evaluations has not been studied.

3. MATERIALS AND PROPOSED METHOD

Patients' pharmaceutical feedback, which were taken from user reviews of drugs, are included in the dataset used in this study [13]. Text messages from 215,063 patients are included in the collection, along with ratings and other information regarding the drugs they were prescribed. Every review has a rate between 1 and 10. The flowchart and steps of the proposed method is shown in Fig. 1.

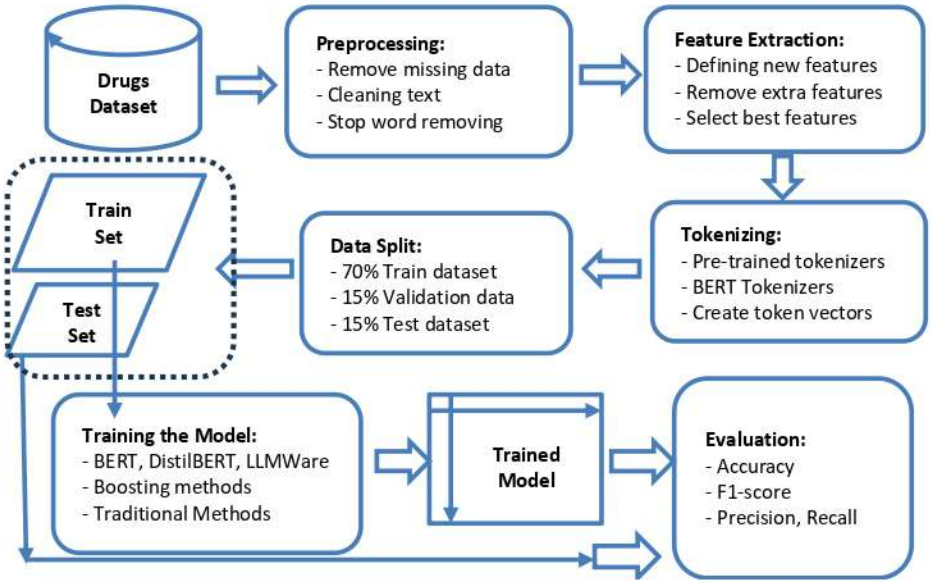

Fig. 1. Flowchart and steps of the proposed method

3.1.

The Medications dataset includes 215,063 records. Each record includes features such as: ID, Drug-name, Condition, Review (text), Rating, etc. Preprocessing tasks are done mostly on the Review feature which shows the

feedback of the patients to medications. The preprocessing on review texts were carried out using the NumPy and NLTK libraries in several steps. First, we remove records with missing data (in any field). Then in the Review texts, the special characters are converted to their correct format (such as punctuation marks). We convert all alphabet characters to lowercase. The resulted reviews are saved as a new text field (for each record) called Cleaned-Review. After removing stop words and stemming the texts using Snowball stemmer the resulted text is stored as a new feature named Cleaned-Review-SS. After fully preprocessing, 213,869 samples remained.

After preprocessing step, we create some extra features from text data. The length of the review, the number of characters in the review, the number of words in the review, day, month, and year of the review. Then we use TextBlob library to obtain the sentiment of each review using Lexicon-based methods. This method creates a number in the range [-1,+1] which represents the polarity of a sentence (-1: negative, +1: positive).

ML models are not able to work directly with texts. So, texts should be converted to vectors of numbers. This study utilizes BERT-based methods to extract tokens from texts. In this way a text is converted to a vector of tokens where each token is represented by a number (an index from a vocabulary).

3.2. Dataset split

Hold-Out cross-validation is used to split the dataset. According to this method, the dataset was randomly divided into 70% training and 30% testing sets. The testing set also is divided to 15% validation and 15% test dataset. The train dataset is used in training the models. The validation dataset is used in the training process for adjusting the best values for the hyper-parameters. Once a model is trained carefully with acceptable accuracy, the model is evaluated using the test set.

3.3. Prediction models

Each sample (record) of the dataset includes a feature called Rating, a value between 1 and 10, which represents the score given to the Review text. Three approaches were implemented in this paper. In the first approach, the rating score were converted into two classes: Negative (for the scores less or equal to 5) and Positive (for the scores greater than 5). In the second approach, the scores were divided into three classes: Negative (for the scores less than or equal to 4), Neutral (for the scores 5 and 6), and Positive (for the scores greater than or equal to 7). Eventually, in the third approach, the dataset scores of this study were divided to five classes (1, 2: Negative; 3, 4: Slightly negative; 5, 6: Neutral; 7, 8: Slightly positive; 9, 10: Positive).

One ML based model, RF, with one gradient boosting method, Light Gradient Boosting Method (LightGBM), and three Large Language Models (LLMs) including BERT [21], DistilBERT [22], and LLMWare [23] were developed to predict patients' sentiments and rate scores. BERT-based models initially are fired with the pre-trained models. Then the models were fine-tuned using the training and validation datasets. The parameters of fine-tuning are represented in the Table 1. In this study, Sklearn and TensorFlow libraries were used for implementation. To determine the ideal hyper-parameter values, Grid Search was used. This method searches and evaluates the hyper-parameters and their values in order to determine the ideal hyper-parameter values for each model.

The best selected hyper-parameters for proposed models are shown in Table 1. We developed our algorithms on a Google Colab with 16GB RAM, T4 GPU with 16GB Memory.

Table 1. The best hyper-parameters selected for the proposed approaches in this study.

Model	Hyperparameters
RF	Max_depth: 30, Criterion: Gini, n_estimators: 200
LGBM	n_estimators=5000, learning_rate=0.10, num_leaves=20, subsample=.8, max_depth=8
DistilBERT	Tokenizer: 'bert-base-uncased', max_len= 128, Optimizer: Adam, Learning_rate=2e-5
LLMWare	Agent: LLMfx, load_tool: Sentiment
BERT	Tokenizer: 'bert-base-uncased', max_len= 128, Optimizer: Adam, Learning_rate=2e-5

The following criteria are used to evaluate the performance of the proposed models:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}, Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN}, F1-Score = \frac{2 \times Precision \times Recall}{Precision+Recall}$$

where TP, TN, FP, and FN are True Positive, True Negative, False Positive, and False Negative, respectively. Moreover, the Area Under Curve (AUC) metric is used to estimate the performance of models using diagrams.

4. RESULTS

One ML model, One boosting based model, and three BERT based models were developed in this paper. This research has considered three different approaches for predicting sentiment classes and rate scores. The first approach considers two classes (positive and negative), The second approach considers three classes (negative, neutral, and positive), and the third approach assumes five classes (negative, slightly negative, neutral, slightly positive, positive). The number of samples of each approach is represented in the Table 2.

Table 2. Distribution of samples in each approach.

Approaches	Class	Number of Samples
First approach	Negative	63,906
	Positive	149,963
Second approach	Negative	53,256
	Neutral	19,053
	Positive	141,560
Third approach	One	37,972
	Two	15,284
	Three	19,053
	Four	37,379
	Five	104,181

The proposed models in the first approach (two classes) were assessed and illustrated in the Table 3. Among traditional ML, boosting based, and BERT-based methods, the highest performance belongs to the BERT, while RF gives promising results and LightGBM giving competitive results among all the methods.

Table 4 represents the assessment of the results of all models in the second approach (three classes). As shown RF again has very promising performance and again LightGBM has verypromising outputs, while BERT has the highest performance among all the models. Table 5 illustrates the results of the models for the third approach (five classes). For the sake of space, we only represent the weighted average of the assessment metrics. As seen, RF and LightGBM having promising results, and BERT has the best performance among all the models.

Table 3. Evaluation of the proposed models in the first approach (two classes).

Model	Class	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
RF	Negative	*89*	*88*	*74*	*80*
	Positive		*90*	*95*	*92*
	Weighted Average		*89*	*89*	*89*
LightGBM	Negative	*89*	*85*	*76*	*80*
	Positive		*90*	*94*	*92*
	Weighted Average		*89*	*89*	*89*
DistillBERT	Negative	90	85	81	83
	Positive		92	94	93
	Weighted Average		90	90	90
LLMWare	Negative	87	74	89	81
	Positive		95	87	91
	Weighted Average		89	87	88
BERT	Negative	***97***	***96***	***96***	***96***
	Positive		***98***	***98***	***98***
	Weighted Average		***98***	***98***	***98***

Table 4. Evaluation of the proposed models in the second approach.

Model	Class	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
RF	Negative	87	85	75	79
	Neutral		100	55	71
	Positive		87	96	91
	Weighted Average		88	87	87
LightGBM	Negative	86	82	75	79
	Neutral		94	54	68
	Positive		87	95	91
	Weighted Average		87	86	86
DistilBERT	Negative	88	85	87	87
	Neutral		44	45	44
	Positive		96	95	95
	Weighted Average		89	88	88
LLMWare	Negative	80	64	88	74
	Neutral		0	0	0
	Positive		93	85	89
	Weighted Average		79	80	79
BERT	Negative	**97**	**97**	**96**	**97**
	Neutral		**87**	**88**	**87**
	Positive		**99**	**99**	**99**
	Weighted Average		**97**	**97**	**97**

Table 5. Evaluation of the proposed models in the third approach.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
RF	78	82	79	79
LightGBM	77	79	78	78
DistilBERT	76	75	76	76
LLMWare	65	54	66	57
BERT	**80**	**81**	**80**	**80**

According to Tables 3-5, the results of the BERT model in all approach, clearly show that this model has outperformed other implemented models. Therefore, the BERT model was utilized and tested for further investigation using general and clinical pre-trained word embeddings. Table 6 represents the comparison with other works.

Table 6. Comparison of the results of this study with previous works on the same dataset.

Study	Method	Classes	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)
[4], 2020	3W3DT-NB	Three classes	88.36	87.35	88.68	88.36
[13], 2018	Logistic regression	Three classes	69.88	-	-	-
[24], 2019	CNN (w2v-entrenable)	Three classes	-	66.72	66.72	66.72
[14], 2019	Deep neural network	Three classes	-	-	84.00	83.00

[15], 2019	Rough-fuzzy feature selection + random forest	Three classes	66.41	-	-	-
This study	Fine-tuned and customized BERT with specific dataset	Two classes	**97.00**	**98.00**	**98.00**	**98.00**
		Three classes	**97.25**	**97.00**	**97.00**	**97.00**
		Five classes	80.35	80.00	81.00	80.00

Fig. 2 and Fig. 3 represent the AUC diagram and confusion matrices of the proposed customized BERT model for approaches of two and three classes, respectively.

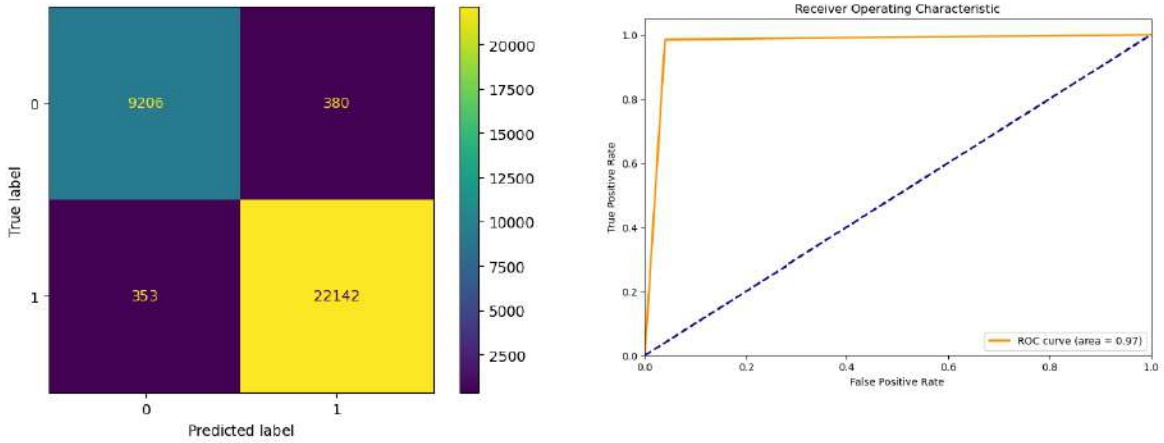

Fig. 2. *Confusion Matrix and AUC of the BERT model for the 2 classes scenario*

5. DISCUSSION

Three distinct situations are used in this study to forecast patients' sentiments toward medicine. In the first case, a rating of five or more was considered positive, while those of less than five were considered negative. In the second case, ratings below five were categorized as negative, five and six are neutral, and those greater than six are set as positive. In the third approach, 1 and 2 are viewed as negative, while 3 and 4 are viewed as slightly negative, and so until 9, and 10 are viewed as positive. The majority of studies employ the first and second approaches as the most popular ways to predict SA [4, 5, 13–15]. In all cases, the accuracy value and F1-score of the best suggested model in Tables 3-6 are higher than those of previous articles. Table 6, Fig. 2, and Fig. 3 show that the suggested best model utilizing the customized BERT model performed well in all approaches.

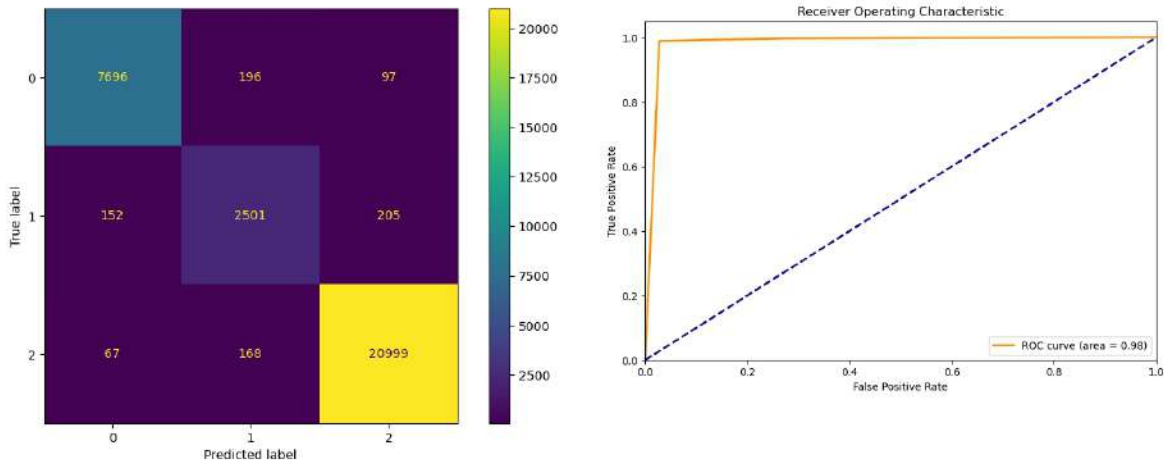

Fig. 3. Confusion Matrix and AUC of the BERT model for the 3 classes scenario

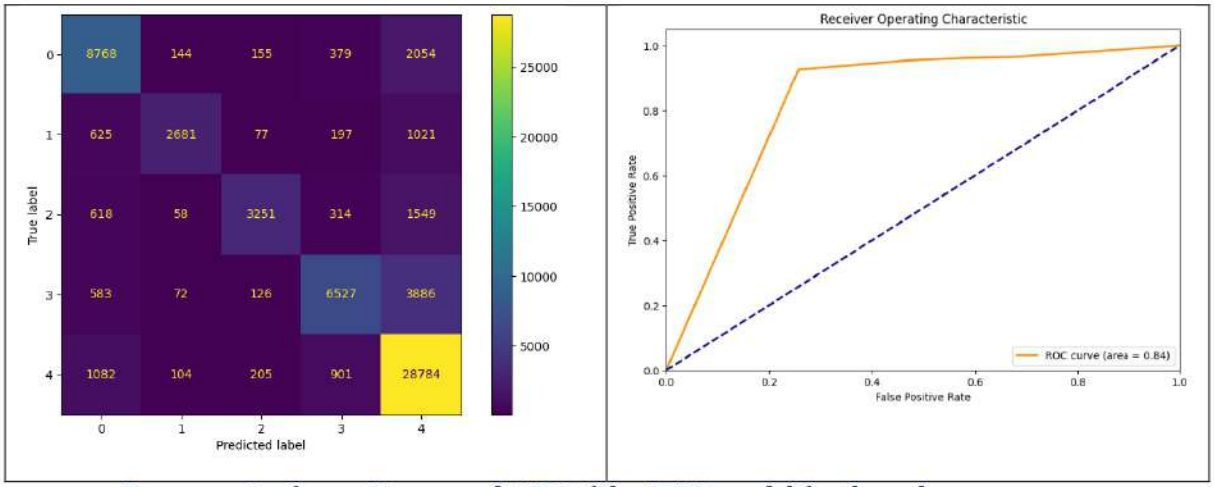

Figure 3: Confusion Matrix and AUC of the BERT model for the 5 classes scenario

By concentrating on SA of patient medication reviews, which posed difficulties necessitating specialized models, this study sought to expand on earlier studies. The performance of the customized BERT model outperformed traditional deep learning and machine learning models when compared to those proposed in earlier studies for medication review datasets, as demonstrated in Table 6.

By combining the advantages of bidirectional transformers, the customized BERT model helps to improve accuracy and robustness while lowering errors. Additionally, the study found that by collecting syntactic and semantic meanings from specific datasets done during fine-tuning the model, performance and generalizability improves. The tailored BERT model in this study, along with adjustments based on particular medical reviews, produced encouraging outcomes in every situation. This study's improvement over previous research was attributed to a number of factors: (1) extensive preprocessing was done on the dataset to reduce bias and errors in the model's predictions; (2) a methodical approach was taken to determine the optimal hyperparameters; and (3) the suggested BERT model for sentiment analysis was used in conjunction with pre-trained parameters in the general domain. These parameters can capture common terminology and language patterns in general text because they have been trained on big datasets. This enhances the model's comprehension and analysis of sentiment in general reviews, resulting in more accurate and contextually relevant predictions in a wider range of situations. They can show good accuracy on the health environment by fine-tuning the model on downstream tasks such as medical texts.

Models such as the suggested customized BERT model may serve as a useful tool for healthcare professionals by assisting in the prediction of drug sentiment. As a plug-in tool, this model can be integrated into software programs for patient medication management. Because the model may be easily integrated into current software programs, neither patients nor clinicians will need to fully comprehend it. This can be a decision support provider for medical staff. It might help doctors prescribe drugs that are more suitable and have fewer adverse effects. Additionally, the results and thorough application of this work can direct the creation of increasingly complex models by medical AI specialists. But using the model in clinical settings raises ethical questions, especially when it comes to using patient-generated data for model deployment and training.

However, this study has many limitations. First, the dataset's imbalance across classes caused variations in the number of cases in each class. Second, there were not enough resources to create more models. Furthermore, no additional high-sample, comparable dataset was available for external validation to evaluate the models' generalizability. Lastly, the customized BERT model has drawbacks that may affect its application in clinical settings, including increased computational requirements and difficulties with interpreting the results. To guarantee efficiency in healthcare settings, these problems can need more improvement.

6. CONCLUSION

The suggested customized BERT model, which was refined using drug data and applied to three approaches, produced very positive outcomes in this investigation. Furthermore, our results demonstrate that pre-trained large language models, fine-tuned using clinical-specific datasets in the clinical domain, outperform other machine learning methods in terms of the models' efficacy.

Furthermore, using the medications dataset, our constructed model outperforms earlier research on SA. We intend to create additional LLM-based family models in the future in order to compare and improve sentiment prediction accuracy.

REFERENCES

[1] Aggarwal, C.C., Zhai, C. (2012). An Introduction to Text Mining. In: Aggarwal, C., Zhai, C. (eds) Mining Text Data. Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-3223-4_1

[2] Opinion Mining. In: Web Data Mining. Data-Centric Systems and Applications. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-37882-2_11, 2007

[3] Cambria, E., Poria, S., Gelbukh, A. & Thelwall, M. Sentiment analysis is a big suitcase. IEEE Intell. Syst. 32, 74–80 (2017).

[4] Basiri, M. E., Abdar, M., Cifci, M. A., Nemati, S. & Acharya, U. R. A novel method for sentiment classification of drug reviews using fusion of deep and machine learning techniques. Knowl. Based Syst. 198, 105949 (2020).

[5] S. M. Qaisar, "Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory," 2020 2nd International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 2020, pp. 1-4, doi: 10.1109/ICCIS49240.2020.9257657.

[6] Asma Ameur, Sana Hamdi, and Sadok Ben Yahia. 2023. Sentiment Analysis for Hotel Reviews: A Systematic Literature Review. ACM Comput. Surv. 56, 2, Article 51 (February 2024), 38 pages. https://doi.org/10.1145/3605152

[7] Chintalapudi, N., Battineni, G., Canio, M. D., Sagaro, G. G. & Amenta, F. Text mining with sentiment analysis on seafarers' medical documents. International Journal of Information Management Data Insights 1, 100005 (2021).

[8] Blanco, A., Casillas, A., Pérez, A. & Diaz de Ilarraza, A. Multi-label clinical document classification: Impact of label-density. Expert Syst. Appl. 138, 112835 (2019).

[9] Denecke, K. & Deng, Y. Sentiment analysis in medical settings: New opportunities and challenges. Artif. Intell. Med. 64, 17–27 (2015).

[10] Chen, J. et al. A classified feature representation three-way decision model for sentiment analysis. Appl. Intell. 52, 7995–8007 (2022).

[11] Gao, Z., Li, Z., Luo, J. & Li, X. Short Text Aspect-Based Sentiment Analysis Based on CNN+ BiGRU. Applied Sciences 12, (2022).

[12] Yu, W., Cui, F. & Hou, Z. The evolution of consumers' demand for hotels under the public health crisis: opinion mining from online reviews. Curr. Issues Tourism 26, 1974–1990 (2023).

[13] Gräßer, F., Kallumadi, S., Malberg, H. & Zaunseder, S. Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. in Proceedings of the 2018 International Conference on Digital Health (ACM, New York, NY, USA, 2018).

[14] Jain, N., Kumar, A., Singh, S., Singh, C. & Tripathi, S. Deceptive reviews detection using deep learning techniques. in Natural Language Processing and Information Systems 79–91 (Springer International Publishing, Cham, 2019).

[15] Chen, T. et al. Sentiment classification of drug reviews using fuzzy-rough feature selection. in 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (IEEE, 2019).

[15] Ebrahimi, M., Yazdavar, A. H., Salim, N. & Eltyeb, S. Recognition of side effects as implicit-opinion words in drug reviews. Online Inf. Rev. 40, 1018–1032 (2016).

[17] Jiménez-Zafra, S. M., Martín-Valdivia, M. T., Molina-González, M. D. & Ureña-López, L. A. How do we talk about doctors and medications? Sentiment analysis in forums expressing opinions for medical domain. Artif. Intell. Med. 93, 50–57 (2019).

[18] Liu, S. & Lee, I. Extracting features with medical sentiment lexicon and position encoding for drug reviews. Health Inf. Sci. Syst. 7, 11 (2019).

[19] Zunic, A., Corcoran, P. & Spasic, I. Sentiment analysis in health and well-being: Systematic review. JMIR Med. Inform. 8, e16023 (2020).

[20] Pilipiec, P., Liwicki, M. & Bota, A. Using machine learning for pharmacovigilance: A systematic review. Pharmaceutics 14, 266 (2022).

[21] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 2019, https://arxiv.org/abs/1810.04805

[22] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. ArXiv. https://arxiv.org/abs/1910.01108

[23] AI for Complex Enterprises, Unified framework for building enterprise RAG pipelines with small, specialized models, https://github.com/llmware-ai/llmware

[24] Colón-Ruiz, C., Segura-Bedmar, I. & Martínez, P. Análisis de Sentimiento en el dominio salud: Analizando comentarios sobre fármacos. Procesamiento del. Lenguaje Nat. 63, 15–22 (2019).