Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 9s (2025)

Using Pre-Trained Distil-Bert Model in Predicting Patients Sentiments Regarding Medical Treatment Results

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 History:

Received: 12-01-2025

Revised: 15-02-2025

Accepted: 01-03-2025

Abstract:

Texts pertaining to medications are among the other fields that has been increased because of information explosion and technological advancements. In medicine, sentiment analysis (SA) is crucial for providing physicians with information about how patients feel about the course of treatment.

This study investigates the application of Large Language Models (LLMs) to predict polarity of patients' opinions (also called sentiment analysis). In this study a dataset including reviews of patients about their satisfaction for treatment and drug prescription is used. Three scenarios are implemented for opinion classification, including two classes (positive, negative), three classes (positive, neutral, negative), and 5 classes (negative, slightly negative, neutral, slightly positive, positive). DistilBERT tokenization method is used for word embedding. For training and fine tuning in clinical domains, three traditional ML based methods, three Boosting based methods, and DistilBERT method, are utilized in model development. We found the best hyper-parameters for all models using Grid-CV search method. The results reveals that the fine-tuned DistilBERT-based model with corresponding word embedding representation, achieved the best results, with accuracy and F1-Score of 90.13% and 90% in two classes, 88.43% and 88% in three classes, and 76.15% and 76% in five classes, respectively. Due to the high accuracy and transparency in decision-making, the proposed models can be used as an auxiliary tool in clinics and medical centers.

Keywords: Machine Learning in Text, Large Language Models in Text Analysis, Explainable Artificial Intelligence, Patients' Opinion Mining, DistilBERT based methods

1. Introduction

The massive volumes of text data generated by various social network, web, and other information-centric applications have drawn more attention to the text mining problem in recent years. The simplest type of data that can be produced for any application scenario is unstructured data. Designing techniques and algorithms that can efficiently handle a broad range of text applications has become imperative [1].

User reviews on websites and social media have significantly increased in the last few decades. On related websites, users post reviews of a wide range of products, including marketing, clinical

services, home appliances, movies, restaurants, and medications. Users of web contribute actively in web-forums, feedback forms, and product review websites. Therefore, modern businesses must make use of this user-generated content to gain more market insights. Therefore, web reviews can be used to determine whether a consumer wants to buy a particular product or if businesses want to know what customers think of their products [2]. Before using or buying goods or services, people read reviews, which helps them make well-informed decisions based on prior evaluations [3-5].

Sentiment Analysis (SA) is a branch of Natural Language Processing (NLP) that investigates to extract the polarity of texts [6]. SA is done in different levels, including aspect, entity, sentence, and document levels [6]. SA is a classification problem, and many AI-based methods are studied to solve this problem [7]. Specifically, many Machine Learning (ML), Deep Learning (DL), and Transfer Learning (TL) methods have been developed to analyze sentiments of users in different fields. For example, sentiment analysis in movie reviews to detect the opinions of users about movies, or in marketplace about the users' opinions on market items, reviews on hotel comments and users' satisfaction on hotels' services, users' opinions on tourist sites, etc., are among the most useful applications of sentiment analysis [8, 9].

Among the most important and practical textual materials are medical and clinical reports, patient opinions, and opinions regarding medical systems and services [7, 10, 11]. Even though the safety of pharmaceuticals is monitored and tested under typical clinical conditions after production, people continue to post their opinions and reviews on websites that review medications. Before taking a medication, people can read reviews about it on these websites. When prescribing and dispensing medication, clinicians and pharmacists can also use this information to consider patients' experiences regarding the efficacy and adverse effects of different medications [7, 10, 11]. Sentiment analysis (SA) of such text data is a helpful step in more precisely assessing the efficacy and adverse effects of medications [10, 11]. Because the outcomes of various medical treatments and awareness of their efficacy have been studied, it is possible to use patient sentiment prediction in medicine to help with future treatment.

AI techniques have been used in numerous studies in the past to identify sentiments in non-clinical settings [3-5, 13-15]. Nevertheless, there are several drawbacks to earlier clinical and medical research [7, 13, 16–17] as follows:

- Limited number of scenarios are considered for predicting sentiments.
- Few studies have examined best hyper-parameter tuning in their study.
- Relatively small datasets are considered in most of the studies in the field.
- No study has compared the various versions of the BERT transformers, as we have done here, on the drug dataset.

This study intends to offer a novel approach for thoroughly determining the sentiments of medication reviews in order to address these limitations. Three Machine Learning, three Boosting Based and one LLM-based models are developed for sentiment and opinion prediction of patients' reviews about drugs. To summarize, the following are our primary contributions in this study:

- We implemented, trained, and compared three ML-based, three Boosting based and one LLM-based models. We tested the models on the drugs dataset, which is a rich dataset for medication reviews.
- We considered three different scenarios to study patients' sentiment classes and ratings.
- Various new preprocessing has 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

Many researchers have tried to examine patient reviews of medications in order to learn more about their needs, feelings, and conditions as well as the negative effects of each drug [7, 13, 16-22]. In order to extract useful information for the aforementioned purpose, SA can play a significant role in this field. Rule-based and conventional ML models have been employed by the majority of researchers for SA [21]. Traditional machine learning algorithms, including Support Vector Machine (SVM), Logistic Regression (LR), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), and K Nearest Neighbor (KNN), were taken into consideration for SA in medication reviews [18, 20, 22, and 23]. Additionally, to enhance the performance of models in SA of medication reviews, new feature extraction techniques and machine learning algorithms have been proposed [23].

To identify side effects in medication reviews, [21] suggested a sentiment extraction recognition system that makes use of SVM and rule-based algorithms. For the purpose of training and testing the model, a medical expert manually selected the medication reviews. According to their findings, the SVM algorithm performed noticeably better than rule-based algorithms. In order to predict the sentiments of overall satisfaction, effectiveness, and side effects of the medications, [18] used LR in both cross-domain and in-domain SA. To get their data, they crawled Druglib.com and Drugs.com. Using data from Drugs.com, they obtained accuracy of 70.06% and kappa of 26.76% in the cross-domain SA. The accuracy and kappa results for the side effects were 49.75% and 25.88%, respectively.

NB, DT, RF, and Ripper have been used to apply SA of the medication reviews to positive, negative, and neutral labels by Chen et al. [20]. They used fuzzy-rough feature selection to efficiently reduce the data in order to address the issue of numerous features. The Term Frequency-Inverse Document Frequency (TF-IDF) and the Bag of Words (BoW) were the two methods they employed to ascertain the terms' values. The accuracy of the RF model was the highest with the value 66.41%.

Lexicon-based and supervised learning SA techniques were used in [22] to inform patients' sentiments for medical and pharmaceutical purposes. Their dataset was in Spanish-language and was

gathered from online discussion boards. According to their findings, it is far more difficult to separate opinions about medications from those about doctors.

For SA of medication reviews, authors in [23] suggested a novel feature extraction method based on position embedding. They used a variety of ML models and feature extraction techniques to classify sentiments and found that their suggested approach performed better.

Authors in [13] predicted how patients will feel about taking each medication based on their medication reviews. They created a CNN for classification with this goal in mind. According to their findings, their method outperformed traditional classification techniques like SVM.

Two distinct approaches have been put forth by Basiri et al. [7] to extract sentiments from patient medication reviews using the conventional ML and DL techniques. They contrasted the DL and conventional models with their best suggested approach and reported that it shows 4% increase in accuracy.

According to these reviews and our study, several gaps are identified in the field of SA in medication. Online reviews of shared medications can be a useful resource for SA, giving useful information to doctors about patients' conditions and drugs' side effects [26]. Prior research has not addressed the development of a unique and thorough sentiment lexicon for medication reviews in the clinical domain [7, 13, 17-23, 27]. Despite the fact that drugs [18] dataset is a rich dataset for SA analysis, no thorough study has made use of it. Moreover, there is no study to employ different pre-trained transformers for SA in medication reviews in the clinical domain and drugs dataset.

3. Materials and Proposed Method

The dataset used in this research contains patients' medication feedback review which was extracted from drugs.com [18]. The dataset contains 215,063 patients' texts along with ratings and some other features about the medication they received. Every review has a score between 1 and 10. The flowchart and steps of the proposed method is illustrated in Fig. 1.

3.1. Data preprocessing

The Drugs 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 drugs. 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.

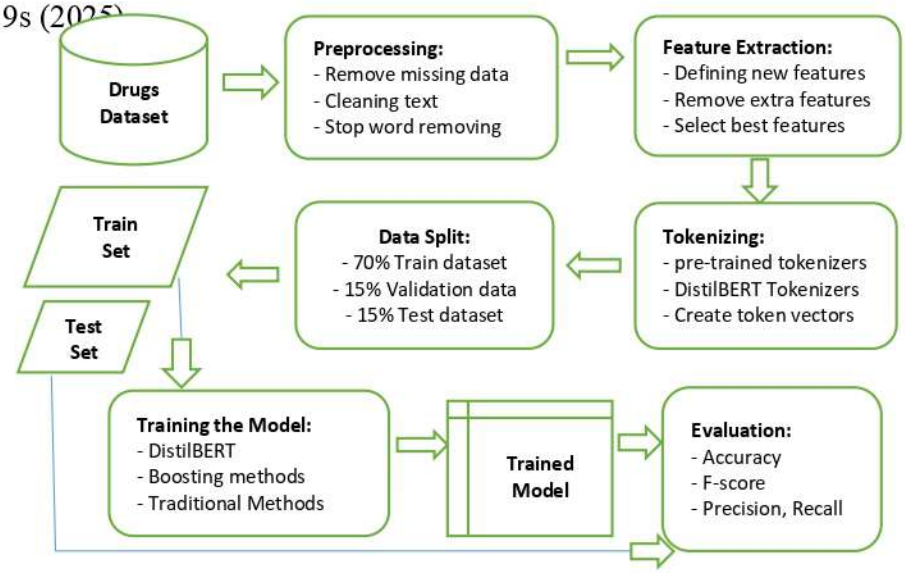

Figure 1: Flowchart and steps of the proposed method

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).

Machine learning models are not able to work directly with texts. So, texts should be converted to vectors of numbers. This study utilizes DistilBERT-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 scenarios were implemented in this paper. In the first scenario, 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 scenario, 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 scenario, the dataset scores of this study were divided to 5 classes (1, 2: Negative; 3, 4: Slightly negative; 5, 6: Neutral; 7, 8: Slightly positive; 9, 10: Positive).

Three machine learning based models, Decision Tree (DT), Random Forest (RF), and Artificial Neural Networks, with three gradient boosting methods including Light Gradient Boosting Method

(LightGBM), Extreme Gradient Boosting Classifier (XGBClassifier), Categorical Boosting Classifier (CatBoostClassifier) and one Bidirectional Encoder Representations from Transformers (BERT-based) [29] model, DistilBERT, were developed to predict patients' sentiments and rate scores. DistilBERT model initially is fired with the pre-trained model. Then the model is fine-tuned using the training and validation datasets. The parameters of fine-tuning is 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 algorithm on a server with Intel Xeon Gold 64-Core-CPU 384 GB of RAM (and google Colab with 16GB RAM, T4 GPU with 16GB Memory).

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

Model	Hyperparameters
DT	Max_depth: 25, Criterion: Gini
RF	Max_depth: 20, Criterion: Gini, n_estimators: 300
ANN	Solver: Adam, Learning_rate_init: 0.1, Hidden_layer_sizes : (20, 10, 1), Alpha: 0.001
LGBM	n_estimators=10000, learning_rate=0.10, num_leaves=30, subsample=.9, max_depth=7
XGBoost	n_estimator = 5000, learning_rate=0.10, num_leaves=40
CatBoost	iterations = 10000, learning_rate = 0.5, num_leaves=30
DistilBERT	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:

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

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

Three ML models, three boosting based models, and a BERT based model were developed in this paper. This research has considered three different scenarios for predicting sentiment classes and rate

scores. The first scenario considers two classes (positive and negative), The second scenario considers three classes (negative, neutral, and positive), and the third scenario assumes 5 classes (negative, slightly negative, neutral, slightly positive, positive). The number of samples of each scenario is represented in the Table 2.

Table 2. Distribution of samples in each scenario.

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

The proposed models in the first scenario (two classes) were assessed and illustrated in the table 3. Among traditional ML, boosted, and DistilBERT method, the highest performance belongs to the DistilBERT, while RF being the best for traditional, and LightGBM being the best among boosting based methods.

Table 4 represents the assessment of the results of all models in the second scenario (three classes). As shown RF again has the best performance among the ML models, LightGBM among the boosted methods, and DistilBERT Has the highest performance among all the models. Table 5 illustrates the results of the models for the third scenario (five classes). For the sake of space, we only represent the weighted average of the assessment metrics. As seen, RF Has the highest performance among ML models, LightGBM among boosted models, and RF has the best performance among all the models.

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

Model	**Class**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
DT	Negative	85	74	76	75
	Positive		90	89	89
	Weighted Average		85	85	85
RF	Negative	*89*	*88*	*74*	*80*

	Positive		*90*	*95*	*92*
	Weighted Average		*89*	*89*	*89*
ANN	Negative	74	62	40	48
	Positive		77	89	83
	Weighted Average		73	74	72
LightGBM	Negative	*89*	*85*	*76*	*80*
	Positive		*90*	*94*	*92*
	Weighted Average		*89*	*89*	*89*
XGBoost	Negative	77	69	42	52
	Positive		79	92	85
	Weighted Average		76	77	76
CatBoost	Negative	88	83	77	80
	Positive		90	93	92
	Weighted Average		88	88	88
DistillBERT	Negative	***90.13***	***85***	***81***	***83***
	Positive		***92***	***94***	***93***
	Weighted Average		***90***	***90***	***90***

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

Model	**Class**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
DT	Negative	81.39	72	74	73
	Neutral		58	61	59
	Positive		88	87	88
	Weighted Average		82	81	81
RF	Negative	87.06	85	75	79
	Neutral		100	55	71
	Positive		87	96	91
	Weighted		88	87	87

	Average				
ANN	Negative	70	55	41	47
	Neutral		0	0	0
	Positive		73	91	81
	Weighted Average		62	70	65
LightGBM	Negative	86.38	82	75	79
	Neutral		94	54	68
	Positive		87	95	91
	Weighted Average		87	86	86
XGBoost	Negative	72.05	62	42	50
	Neutral		80	0	0
	Positive		74	93	83
	Weighted Average		72	72	67
CatBoost	Negative	86.13	80	77	78
	Neutral		84	57	68
	Positive		88	94	91
	Weighted Average		86	86	86
DistilBERT	Negative	**88.43**	**85**	**87**	**87**
	Neutral		**44**	**45**	**44**
	Positive		**96**	**95**	**95**
	Weighted Average		**89**	**88**	**88**

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

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
DT	72.31	72	72	72
RF	**78.92**	**82**	**79**	**79**

ANN	52.40	32	54	41
LightGBM	77.94	**79**	**78**	**78**
XGBoost	54.30	48	54	43
CatBoost	76.87	77	77	77
DistilBERT	76.15	75	76	76

According to Tables 3-5, the results of the DistilBERT model in first and second scenarios, clearly show that this model has outperformed other implemented models. Therefore, the DistilBERT 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 (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
[7], 2020	3W3DT-NB	Three classes	88.36	88.68	88.36	87.35
[32], 2019	CNN (w2v-entrenable)	Three classes	-	66.72	66.72	66.72
[18], 2018	Logistic regression	Three classes	69.88	-	-	-
[19], 2019	Deep neural network	Three classes	-	84.00	83.00	-
[20], 2019	Fuzzy-rough feature selection + Random forest	Three classes	66.41	-	-	-
This study	DistilBERT (Two and Three classes) (RF: five classes)	Two classes	**90.13**	**90**	**90**	**90**
		Three classes	**88.43**	**89**	**88**	**88**
		Five classes	**78.92**	**82**	**79**	**79**

Figures 2 and 3 represent the AUC diagram and confusion matrices of the proposed DistilBERT model for scenarios of two and three classes, respectively.

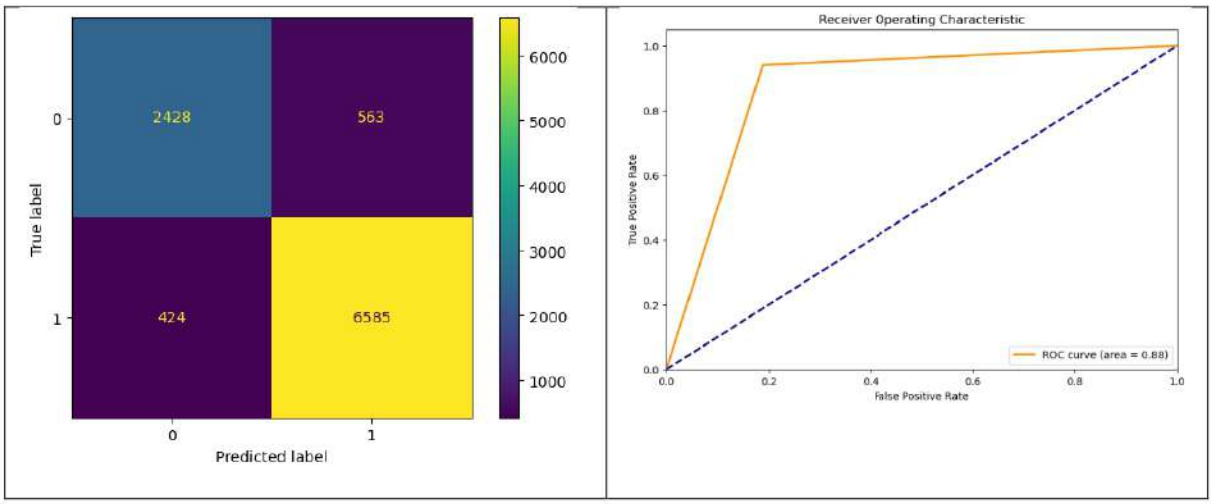

Figure 2: Confusion Matrix and AUC of DistilBERT for the 2 classes scenario

5. Discussion

Three distinct situations are used in this study to forecast patients' SA toward medicine. The majority of studies employ the first and second scenarios as the most popular ways to predict SA [7, 8, 18–20]. In the first case, a score of five or more was considered positive, while a score of less than five was considered bad.

In the second case, scores below five were categorized as negative, scores between five and six as neutral, and scores greater than six as positive. In the third scenario, 1 and 2 are viewed as negative, while 3 and 4 are viewed as somewhat unfavorable till 9, and 10 are viewed as positive.

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. According to [32], the findings shown in Figures 4 and 5 show that the suggested best model utilizing the DistilBERT model performed admirably in the first and second scenarios and passably in the third.

By concentrating on SA of patient medication reviews, which posed difficulties necessitating specialized models, this study sought to expand on earlier studies.

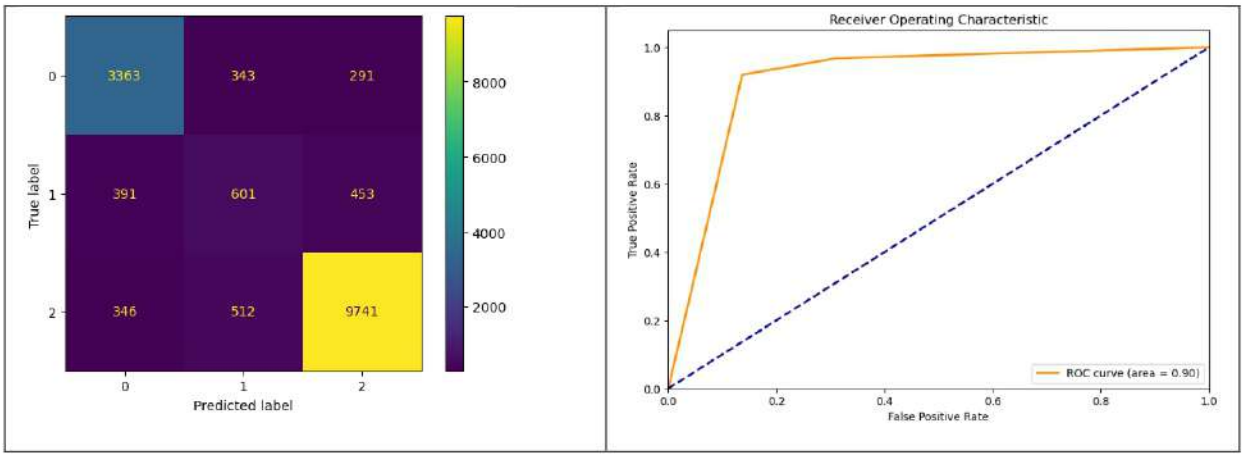

Figure 3: Confusion Matrix and AUC of DistilBERT for the 3 classes scenario

6. Conclusion

The suggested DistilBERT model, which was refined using drug data and applied to three scenarios, produced very positive outcomes in this investigation. Furthermore, our results demonstrate that pre-trained word embedding in the clinical domain outperforms pre-trained word embedding in the general domain in terms of the models' performance.

Furthermore, using the drugs.com dataset, our constructed model outperforms earlier research on SA. We intend to create additional BERT-based family models in the future in order to compare and improve SA's 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., Wang, H. & White, B. Guest editorial: Big social data analysis. Knowl. Based Syst. 69, 1–2 (2014).

[4] Budennyy, S., Kazakov, A., Kovtun, E. & Zhukov, L. New drugs and stock market: a machine learning framework for predicting pharma market reaction to clinical trial announcements. Sci. Rep. 13, 12817 (2023).

[5] Grassi, M., Cambria, E., Hussain, A. & Piazza, F. Sentic web: A new paradigm for managing social media affective information. Cognit. Comput. 3, 480–489 (2011).

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

[7] 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).

[8] 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.

[9] 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

[10] 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).

[11] 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).

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

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

[14] Wu, D. C., Zhong, S., Qiu, R. T. R. & Wu, J. Are customer reviews just reviews? Hotel forecasting using sentiment analysis. Tourism Econ. 28, 795–816 (2022).

[15] Sarsam, S. M., Al-Samarraie, H., Alzahrani, A. I., Alnumay, W. & Smith, A. P. A lexicon-based approach to detecting suicide-related messages on Twitter. Biomed. Signal Process. Control 65, 102355 (2021).

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

[17] 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).

[18] 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).

[19] 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).

[20] 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).

[21] 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).

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

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

[24] Xu, Q. A., Chang, V. & Jayne, C. A systematic review of social media-based sentiment analysis: Emerging trends and challenges. Decision Analytics Journal 3, 100073 (2022).

[25] Cascini, F. et al. Social media and attitudes towards a COVID-19 vaccination: A systematic review of the literature. EClinicalMedicine 48, 101454 (2022).

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

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

[28] Zhang, Y., Jin, R. & Zhou, Z.-H. Understanding bag-of-words model: a statistical framework. Int. J. Mach. Learn. Cybern. 1, 43–52 (2010).

[29] 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

[30] Hossin, M. & Sulaiman, M. N. A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process 5, (2015).

[31] Linardatos, P., Papastefanopoulos, V. & Kotsiantis, S. Explainable AI: A review of machine learning interpretability methods. Entropy (Basel) 23, 18 (2020).

[32] 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).