2023 14th International Conference on Intelligent Systems: Theories and Applications (SITA) | 979-8-3503-0821-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/SITA60746.2023.10373604

Adverse Drug Reactions Detection from Social Media: an Empirical Evaluation of Machine Learning Techniques

Oumayma ELBIACH
LISAC Laboratory, Faculty of
Sciences Dhar El Mahraz,
Sidi Mohamed Ben Abdellah University
Fez, Morocco
oumayma.elbiach@usmba.ac.ma

Hanane GRISSETTE
MIMSC Laboratory, Higher
School of Technology,
Cadi Ayyad University
Marrakech, Morocco
hanane.grissette@usmba.ac.ma

El Habib NFAOUI
LISAC Laboratory, Faculty of
Sciences Dhar El Mahraz,
Sidi Mohamed Ben Abdellah University
Fez, Morocco
elhabib.nfaoui@usmba.ac.ma

***Abstract*—This research addresses the critical concern of Adverse Drug Reactions (ADRs), emphasizing the need for their timely detection to safeguard patient well-being. Detecting ADRs within sentences is pivotal for effective public health monitoring. The study's primary objective is to assess whether sentences contain ADR references, a crucial step in identifying potential ADRs early. Timely recognition can mitigate patient harm and enhance drug development processes. The rise of patient engagement on social media has turned it into a valuable real-time resource for ADR-related information. Patients increasingly share their personal narratives about drug usage on these platforms, creating an innovative avenue for gathering firsthand accounts of ADRs. This evolving trend has transformed social media into an indispensable source of information. The study aims to compare machine learning algorithms in classifying sentences as containing ADRs or not. Three diverse datasets—CADEC, TwiMed (PubMed), and ADE—are used to train and evaluate models. Rigorous experimentation highlights the superiority of the Naive Bayes classifier over other methods. Notably, this classifier achieves remarkable accuracy rates of 94.29%, 78.76%, and 64.93%, on the CADEC, ADE, and PubMed datasets, respectively. This comparative study demonstrates the effectiveness of machine learning in identifying ADRs within sentences and underscores the Naive Bayes classifier's consistently impressive performance across different datasets.**

***Index Terms*—adverse drug reaction, social media, machine learning, text classification, natural language processing**

I. Introduction

Adverse drug reactions (ADRs) are a significant public health issue. With the development of new drugs, drug safety issues, particularly ADRs, have become more prominent [1]. The detection of adverse drug reaction entities from texts is a critical task for the public health monitoring process, it aims to automatically determine whether a sentence contains an ADR or not, which is a fundamental study for pharmacovigilance. Every drug has advantages, but it does not always produce the desired effect for its users. Due to clinical trial limits in terms of scale and time, it is hard to conduct a comprehensive evaluation of the results of a specific drug before it is released to the market. According to current statistics, ADRs cause irreversible healthcare harm to the public.

In early studies, clinical electronic medical records and the Federal Drug Administration's Adverse Event Reporting System (FAERS) are significant spontaneous reporting systems [2]. But, they suffer from updating regularly and timely their reports. An alternative approach to identify ADRs in a timely manner on a broader scale is to use social media. Internet and social media have become an integral part of people's daily life. With the prosperity of social media, it is increasingly being used to share with people who have similar health concerns and to exchange information about their health problems, their treatment experiences, and post their use of prescription drugs. Because of this behavior, user posts on social media are an important source of ADR-related information and are more real-time.

Therefore, ADR detection and the use of statistical data can help doctors in reducing clinical risks associated with ADRs, as well as lowering healthcare costs for society when prescribing drugs [3]. The timely detection of ADRs is crucial and depends on an efficient ADR-reporting process. However, the detection of ADRs from social media has two main challenges: 1) The language on social media is informal, with various colloquialisms used in descriptions. It includes abbreviations, misspellings, and phrase construction irregularities that make extraction more difficult. 2) There are small annotated corpora, particularly for social media data. These challenges introduce various levels of noise to ADR signals that can be captured from social media.

Machine learning offers a promising approach to classify adverse drug reactions (ADRs) by leveraging a patient's symptoms, medical history, and drug usage. The ultimate goal is to identify potential ADRs at an early stage, enhancing patient safety. In this process, machine learning text classification plays a crucial role. Models are trained using historical data, enabling them to make precise predictions. By supplying prelabeled examples as training data, machine learning algorithms learn to recognize patterns and relationships between input

979-8-3503-0821-1/23/$31.00 ©2023 IEEE

data and their corresponding outputs, thereby improving the accuracy of ADR classification. This approach holds significant potential in advancing healthcare and ensuring timely intervention to prevent severe ADRs.

In our research, we conduct a comparative analysis using different machine learning techniques to detect adverse drug reactions. The first hurdle we encountered was the imbalanced data distribution in the TwiMed (Pubmed), CADEC, and ADE datasets. To overcome this challenge, we adopted two approaches. Firstly, we applied SMOTE (Synthetic Minority Over-sampling Technique) for oversampling to balance the datasets of CADEC and TwiMed (PubMed) during the model training phase. Secondly, for the ADE dataset, where there was a significant disparity between positive and negative classes, we implemented undersampling. This approach effectively addressed the data constraints, leading to promising results in our study. By handling the imbalanced data problem, our research provides more reliable and accurate insights into adverse drug reaction detection. The remainder of this paper is structured as follows: Section 2 offers an overview of existing methods for ADR detection. Section 3 outlines the machine learning approach employed for ADR classification. In Section 4, we present the datasets used, the experimental results, and a discussion. Lastly, Section 5 concludes the paper.

II. Related Work

As social media networks continue to play an increasingly important role in daily life and behavior, they become a preferred platform for sharing health information and discussing various drug-related issues. In light of this, researchers have turned to these platforms to identify Adverse Drug Reactions (ADRs), which have become a significant public health concern in various communities. Previously, researchers primarily relied on lexicon-based approaches [4] to identify Adverse Drug Reactions (ADRs) in text. They tended to utilize unsupervised methods for statistical analysis [[5], [6], [7]. Lexicon-based approaches have limitations in their ability to detect expressions that are not included in the lexicons, resulting in a recall that is impacted by these limitations.

Traditional machine learning methods have become more prevalent in ADR detection with the emergence of annotated data. [8] employed Decision Trees, Maximum Entropy, and SVMs with a large number of engineered features for ADR detection. They achieved an F-score of 77% for the ADR class using the ADE dataset. [9] developed a machine learning model to extract adverse drug reactions (ADRs) from MEDLINE case reports. To annotate the reports, they employed an ontology-driven methodology. Remarkably, their model achieved an impressive F-score of 87%, demonstrating its effectiveness in accurately identifying ADRs from the text data. In the study conducted by [10], the authors used NLP techniques to extract rich features from text to improve binary classification performance using three supervised classification approaches: Naïve Bayes (NB), Support Vector Machines (SVM), and Maximum Entropy (ME). The researchers utilized three datasets, namely ADE, TW, and DS. Among the three classification methods, SVM outperformed the others, achieving ADR F-scores of 81.2%, 53.8%, and 67.8%, respectively. In their study, [11] focused on ADR classification using a set of features for binary classification. Their feature-rich classifier employed Linear SVM and Logistic Regression algorithms. Notably, the best results were achieved with SVM, reaching an accuracy of 80.3% on the CADEC dataset. On the other hand, Logistic Regression achieved an accuracy of 73.3% on the Twitter dataset. [12] introduced a novel method that can extract deep linguistic features and then combine them with shallow linguistic features for ADR detection. As a result of their approach, remarkable AUCs of 94.44% and 88.97% were attained when evaluating the method on the DailyStrength and Twitter datasets, respectively. [13] presents an Imdb movie review text classification model, including three phases: data preprocessing, text weighting, and classifier development. Evaluation involves comparing various classification algorithms using accuracy metrics. Logistic Regression with Bi-grams and Support Vector Machine with tf-idf achieved the highest accuracy in an 80:20 data split scenario, while the combination of Tf-idf with Bi-grams performed the worst. [14] reviews clinical recommendation systems and ADR categorization using various models, with deep neural networks (DNN) showing the best ADR detection performance. SVM exhibits significant improvement post-preprocessing and with clinical vector space integration, achieving 86% accuracy in ADR classification with the SIDER dataset.

III. Adverse Drug Reaction Detection

A. A general framework for ADR detection based on machine learning

The problem of adverse drug reaction detection, utilizing machine learning algorithms, revolves around creating a model capable of identifying and categorizing adverse reactions caused by medications, relying on patient data and drug-related information. This is typically accomplished through the utilization of supervised learning techniques, where the model is trained on labeled data to learn patterns and make predictions about adverse reactions.

The problem can be formulated as a binary classification task, where the model is trained to classify whether a patient has experienced an adverse reaction or not. The input data for the model typically includes patient data, medical history, medication history, and any reported symptoms or side effects. The output of the model is a binary prediction indicating the presence or absence of an adverse reaction.

To develop an effective model, as depicted in the figure 1, the following steps are typically taken:

- Data collection: Collecting a large, diverse dataset of patient information, drug information, and adverse reaction labels.
- Data preprocessing: Cleaning and preprocessing the data to ensure it is accurate, consistent, and ready for machine learning.
- Feature engineering: Selecting or creating relevant features that capture important information about the patient,

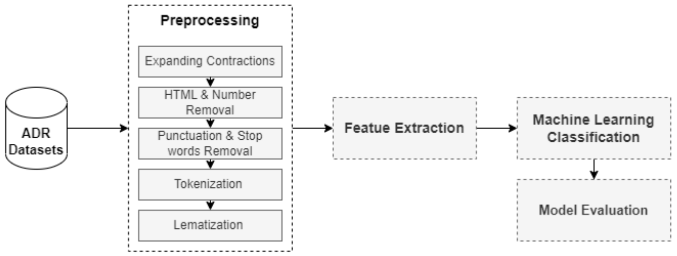

Fig. 1. An illustration of ADR detection methodology.

medication, and potential adverse reactions. Natural language processing techniques can also be used to extract features from textual information, such as medical notes.

- Model selection: Choosing an appropriate machine learning algorithm that is well-suited to the problem and the data.
- Model training: Using the prepared data to train the model, adjusting its parameters to optimize performance.
- Model evaluation: Testing the model on a separate test dataset to evaluate its performance, measuring metrics such as accuracy, precision, recall, F1 score, and ROC AUC.

To gain a thorough understanding of our study's approach, we will begin by presenting the mathematical formulation of the adverse drug reaction problem, utilizing machine learning algorithms. Subsequently, we will comprehensively explain all the algorithms employed in the context of adverse drug reaction detection.

B. Adverse Drug Reaction Detection based on machine learning techniques

The adverse drug reactions detection problem in machine learning can be formulated as a classification problem, where the goal is to learn a mapping from a set of features to a binary label indicating whether an adverse drug reaction occurred or not.

Let $X = x_1, x_2, ..., x_n$ be a set of n medical records, where each record consists of a set of m features describing the patient, their medical history, and the drug they were prescribed. The features can be represented as a matrix X, where each row represents a record and each column represents a feature.

Let $Y = y_1, y_2, ..., y_n$ be a set of binary labels indicating whether an adverse drug reaction occurred or not for each medical record in X. The labels can be represented as a vector Y of length n.

The goal of the machine learning algorithm is to learn a function $f(X)$ that maps the input features X to the output labels Y. This function can be represented as a hypothesis $h(X)$, which is parameterized by a set of weights w. The weights w are learned by minimizing a loss function $L(w)$ that measures the discrepancy between the predicted labels and the true labels.

One commonly used loss function for binary classification problems is the binary cross-entropy loss, which can be defined as:

$$L(w) = -1/n * \sum (y_i * log(h(x_i)) + (1 - y_i) * log(1 - h(x_i))) \tag{1}$$

where $h(x_i)$ represents the predicted probability of an adverse drug reaction for the $i-th$ medical record and y_i is the true label for the $i-th$ medical record.

The problem can be solved using various machine learning algorithms, such as support vector classification, Decision tree, random forests, logistic regression, XGBoost, multinomial naive Bayes, AdaBoost, bagging, and voting. The choice of algorithm depends on the nature of the data and the complexity of the problem.

Once the model is trained, it can be used to predict the occurrence of adverse drug reactions for new patients based on their medical history and the drugs they are prescribed.

In this part, we will provide a brief and succinct overview of the machine learning algorithm that we have employed in our ADR detection methodology.

1) Support Vector Classification (SVC): Support Vector Classification is a supervised machine learning algorithm used for binary classification tasks. It aims to find the optimal hyperplane that best separates two classes in a high-dimensional feature space.

2) Decision Tree Algorithm (DT): Decision Tree is a supervised machine learning algorithm used for both classification and regression tasks. It recursively splits the data based on features to create a hierarchical tree, where each leaf node represents a class label (for classification) or a predicted value (for regression). It is interpretable, capable of handling nonlinear relationships, and can be prone to overfitting on complex datasets.

3) Random Forest Algorithm (RF): Random Forest is an ensemble learning method that combines multiple decision trees to improve accuracy and reduce overfitting. It creates each tree using a random subset of features and data and then aggregates their predictions to make the final classification or regression decision.

4) Logitic Regression Algorithm (LR): Logistic Regression is a popular binary classification algorithm that models the probability of an instance belonging to a particular class using a logistic function. It estimates the coefficients of input features to fit a decision boundary and makes predictions based on the probability threshold of 0.5. It's widely used due to its simplicity, interpretability, and efficiency for linearly separable data.

5) XGBoost algorithm (XGB): XGBoost is a powerful gradient-boosting algorithm known for its high performance in both regression and classification tasks. It uses an ensemble of weak decision tree learners, iteratively refining predictions by minimizing a loss function and employing regularization techniques. XGBoost is favored for its speed, scalability, and ability to handle complex, large-scale datasets with exceptional accuracy.

6) Multinomila Naive Bayes Algorithm (Multinomial NB): Multinomial Naive Bayes is a probabilistic classification algorithm that extends the Naive Bayes method for handling discrete feature data, commonly used for text classification tasks where features represent word counts or frequencies.

7) Adaptive Boosting Algorithm (AdaBoost): Adaptive Boosting is an ensemble learning algorithm that iteratively combines multiple weak learners (typically decision trees) to create a strong classifier. It assigns higher weights to misclassified instances in each iteration, focusing on the difficult cases.

8) Bagging Algorithm: Bagging (Bootstrap Aggregating) is an ensemble learning technique that combines predictions from multiple models trained on different subsets of the data to improve prediction accuracy and reduce variance. It is particularly useful for high-variance models, enhancing overall performance and stability.

9) Voting Algorithm: Voting is an ensemble learning technique that combines predictions from multiple models to make the final decision based on a majority vote (for classification) or average (for regression), leading to improved prediction accuracy and robustness.

IV. Experiments and Results

In this section, we present the experimental results and evaluate the effectiveness of 9 machine learning algorithms that we have employed in our ADR detection methodology.

A. Datasets and Settings

We conduct an empirical evaluation of the suggested algorithms using three datasets: TwiMed, CADEC, and ADE. Table 1 displays summary statistics for corpora.

TABLE I
Summary statistics of the corpora.

Corpus	Documents	ADR	No ADR
TwiMed(PubMed)	1000	148	852
CADEC	1186	1051	135
ADE	1644	6821	16695

- **TwiMed** [15]: TwiMed corpus consists of two parts: TwiMed-PubMed and TwiMed-Twitter, which are the sentences that were collected from PubMed and Twitter, respectively. In this study, we used only TwiMed (PubMed) which contains 1000 PubMed sentences. It consists of three types of entities: Drugs, Diseases, and Symptoms. Furthermore, it includes three kinds of relationships between those entities: reason-to-use, outcome-positive, and outcome-negative. Symptoms and Diseases are both considered adverse reactions in our experiments. The term outcome-negative refers to the possibility that the drugs in the sentence could cause adverse reactions. We labeled the sentence as ADR if the relationship between adverse reactions and drugs was identified as Outcome-negative in the sentence; otherwise, we annotated it as No ADR.
- **CSIRO Adverse Drug Event Corpus (CAEDC)** [16]: The CADEC dataset was sourced from social media posts, where the sentences predominantly employ colloquial language and diverge from conventional punctuation and formal English grammar. The dataset encompasses five types of entities: ADR, Drug, Disease, Symptom, and Finding. For our specific study, our primary focus lies on the ADR entity. Therefore, sentences containing an ADR were labeled as ”ADR,” while those without any ADR mention were labeled as ”No ADR”.
- **ADE** [17]: The ADE corpus is sourced from 1644 PubMed abstracts, containing sentences that indicate the presence or absence of Adverse Drug Reactions (ADRs) obtained from medical case reports. Among these sentences, 6821 are labeled as positive instances (containing at least one ADE), while 16,695 sentences are labeled as negative instances (without any ADE). The dataset has been carefully divided into these two categories to facilitate detailed analysis and support ADR classification tasks effectively.

B. Data Preparation for ADR detection

Data preprocessing is an important first step before using any classification algorithms since algorithms learn from data and the effectiveness of learning for issue solving depends on the relevant data necessary to solve a given problem, known as features.

Data Cleaning: Data cleansing involves identifying and removing irrelevant and unnecessary data. To clean our data, we employed various techniques, including expanding contractions, removal of HTML tags and numerical characters, elimination of punctuation marks and stop words, tokenization, and lemmatization. The following figure **??** illustrates the cleaning process applied to an example text extracted from the CADEC dataset.

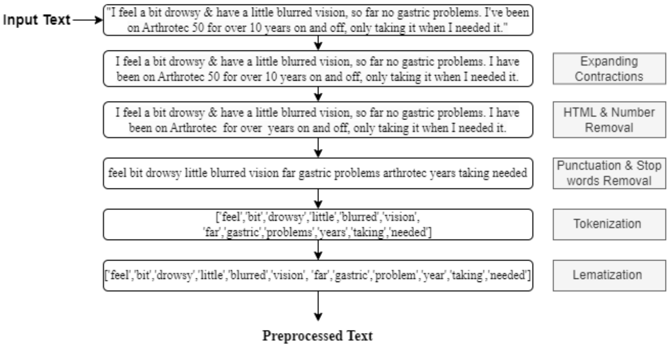
Input Text
"I feel a bit drowsy & have a little blurred vision, so far no gastric problems. I've been on Arthrotec 50 for over 10 years on and off, only taking it when I needed it."
I feel a bit drowsy & have a little blurred vision, so far no gastric problems. I have been on Arthrotec 50 for over 10 years on and off, only taking it when I needed it.
Expanding Contractions
I feel a bit drowsy & have a little blurred vision, so far no gastric problems. I have been on Arthrotec for over years on and off, only taking it when I needed it.
HTML & Number Removal
feel bit drowsy little blurred vision far gastric problems arthrotec years taking needed
Punctuation & Stop words Removal
['feel','bit','drowsy','little','blurred','vision', 'far','gastric','problems','years','taking','needed']
Tokenization
['feel','bit','drowsy','little','blurred','vision', 'far','gastric','problem','year','taking','needed']
Lematization
Preprocessed Text

Fig. 2. Visualization of the ADR Preprocessing Stage.

Data Balancing: Data balancing is vital in machine learning algorithms to avoid biased and inaccurate models caused by imbalanced datasets. Skewed class distributions lead to poor performance in minority classes, favoring the majority class. Oversampling SMOTE (Synthetic Minority Over-sampling

Technique) [18] addresses this issue by generating synthetic samples for the minority class, ensuring fair representation for all classes. SMOTE improves generalization and accuracy on underrepresented classes while reducing the risk of bias. Integrating SMOTE during data preprocessing enhances the fairness and effectiveness of machine learning algorithms, promoting more equitable and inclusive decision-making in various applications. In our study, we applied SMOTE to CADEC and PubmMed datasets, but not to ADE due to unsatisfactory results. Instead, we employed undersampling on the majority class and removed duplicate ADR sentences, resulting in 2865 samples for each class.

Feature Extraction: Before applying machine learning algorithms, it is necessary to convert data into a format that a machine can understand, which is known as feature engineering. In this investigation, we applied the inverse document frequency (TF-IDF) approach and the word frequency. It is used to assess which words of a corpus are likely to be favorable based on their document frequency. Words with higher TF-IDF values are considered to have a stronger relationship within the document in which they appear. The TF-IDF formula is as follows:

$$TF = \frac{\text{Number of times a word 'X' appears in a Document}}{\text{Number of words present in a Document}} \tag{2}$$

$$IDF = \log\left(\frac{\text{Number of Document present in a corpus}}{\text{Number of Documents where word 'X' has appeared}}\right) \tag{3}$$

$$TF - IDF = TF * IDF \tag{4}$$

C. Data Splitting

In accordance with best practices in data preparation for our research, we performed a standard data splitting procedure, allocating 70% of the dataset for training and reserving the remaining 30% for testing. This approach ensures a rigorous evaluation of our model's performance on unseen data, enhancing the reliability and generalization of our findings, which we believe is crucial for presenting robust results.

D. Evaluation metrics

Applying evaluation metrics is necessary to evaluate the effectiveness of models and identify the best model based on these metrics in order to assess the model's performance. In this study, we used Precision, Recall, Accuracy, and F1-score, ROC AUC. Generally, Precision measures the proportion of correctly identified positive instances out of all predicted positive instances. Recall measures the proportion of correctly identified positive instances out of all actual positive instances. Accuracy measures the percentage of correctly classified instances in the dataset. F1 Score is the harmonic mean of precision and recall and is used to measure the overall performance of a classifier. ROC AUC (Area Under the Receiver Operating Characteristic Curve) is a metric that quantifies the ability of a binary classification model to distinguish between positive and negative classes.

V. Results and Discussion

The following table II presents a summary of the results obtained from several evaluation metrics including accuracy, precision, recall, and F1 score, obtained from different machine learning classifiers. Additionally, Figure 2 showcases the results of the ROC AUC metric, providing a comprehensive view of the classifiers' performance.

TABLE II
Performance Comparison of Various ML Algorithms on Three Datasets.

PubMed Dataset				
ML Classifier	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Multinomial NB	91%	65.78%	64.10%	64.93%
SVC	91%	70%	53.84%	60.86%
LR	90.33%	70.83%	43.58%	53.96%
Voting	90%	69.56%	41.02%	51.61%
XGB	89.66%	65.38%	43.58%	52.30%
RF	86.33%	86.33%	30.76%	36.92%
Bagging	85%	42.85%	46.15%	41.02%
AdaBoost	84.66%	43.63%	61.53%	51.06%
DT	84.33%	41.30%	48.71%	44.70%
CADEC Dataset				
Multinomial NB	89.32%	89.71%	99.36%	94.29%
Voting	89.04%	89.68%	99.05%	94.13%
SVC	89.04%	89.91%	98.73%	94.11%
LR	88.76%	89.65%	98.73%	93.97%
XGB	88.48%	90.08%	97.78%	93.77%
RF	88.20%	88.70%	99.36%	93.73%
Bagging	87.35%	90.20%	96.20%	93.10%
AdaBoost	86.51%	90.85%	94.30%	92.54%
DT	82.30%	82.30%	90.18%	90.04%
ADE Dataset				
SVC	79.05%	80.51%	74.90%	77.61%
Voting	78.35%	78.77%	75.75%	77.23%
LR	78.18%	78.55%	75.63%	77.06%
Multinomial NB	77.54%	72.69%	85.95%	78.76%
RF	74.86%	73.95%	74.30%	74.13%
XGB	74.63%	77.15%	67.70%	72.12%
Bagging	74.22%	73.55%	73.10%	73.32%
DT	71.02%	69.09%	72.74%	70.87%
AdaBoost	69.86%	77.68%	53.06%	63.05%

Table II clearly illustrates the outstanding performance of the naive Bayes algorithm in classifying adverse drug reactions (ADRs) across three datasets. This remarkable proficiency highlights its aptitude for managing textual data, showcasing the algorithm's effectiveness in capturing subtle linguistic patterns associated with ADRs, owing to its utilization of a probabilistic framework and the assumption of feature independence. Naive Bayes can perform well with relatively small datasets, which is often the case in ADR classification where labeled data can be scarce. Notably, it achieved remarkable F1 values of 64.93%, 94.29%, and 78.76% in the PubMed, CADEC, and ADE datasets, respectively. The Adaboost algorithm stands out by achieving the highest precision of 91.30% in the CADEC dataset. This outstanding precision underscores the algorithm's ability to correctly identify positive cases, minimizing false positives and improving the overall precision of the classification. Both the Naive Bayes and Random Forest algorithms demonstrated good performance on the CADEC dataset, achieving the highest recall value of 99.36%. This outstanding recall indicates their ability to effectively identify a vast majority of positive cases correctly. In contrast to the promising performance of other algorithms, bagging, decision tree, and AdaBoost demonstrate relatively poor performance on the PubMed, CADEC, and ADE datasets, respectively. These algorithms did not yield competitive results compared to their counterparts, indicating that they may not be well-suited for handling the specific characteristics of the datasets.

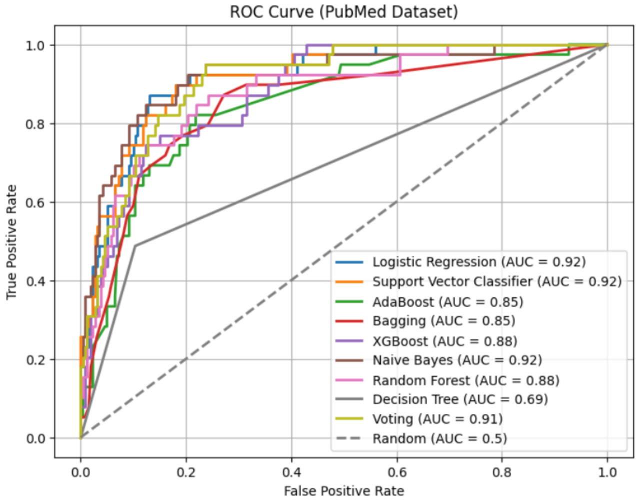

Fig. 3. ROC Evaluation: Assessing the Performance of Different ML Algorithms on PubMed Dataset.

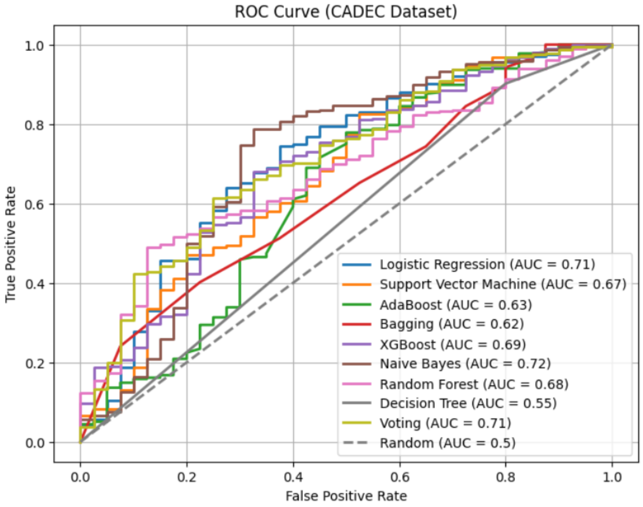

Fig. 4. ROC Evaluation: Assessing the Performance of Different ML Algorithms on CADEC Dataset.

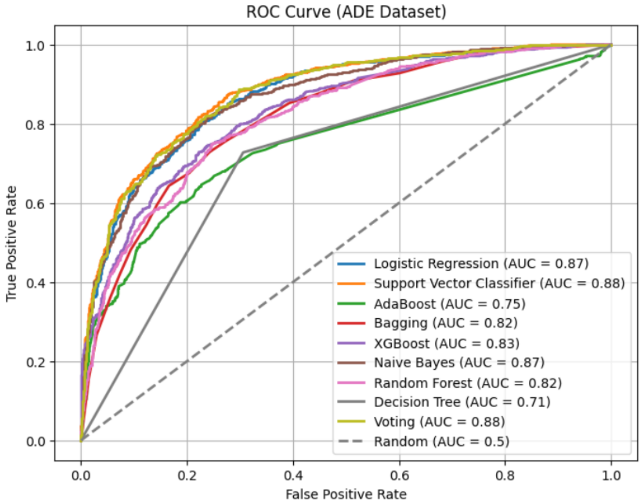

Fig. 5. ROC Evaluation: Assessing the Performance of Different ML Algorithms on ADE Dataset.

Furthermore, the ROC values for the three datasets are depicted in Figures 3, 4, and 5. When analyzing the CADEC dataset, it becomes evident that both the Naive Bayes and Voting algorithms stand out by achieving a notable ROC value of 0.72. Turning our attention to the PubMed dataset, the Logistic Regression, Support Vector Classifier (SVC), and Naive Bayes algorithms display commendable performance, achieving an impressive ROC value of 0.92. Shifting focus to the ADE dataset, the Support Vector Classifier (SVC) algorithm takes the lead with the highest ROC value of 0.88. These outcomes underscore the distinct performance dynamics of the algorithms across the diverse datasets, demonstrating their sensitivity to dataset characteristics.

VI. Conclusion

Adverse drug reactions (ADRs) refer to the negative side effects caused by prescribed medication. These reactions can vary by age group, and it is important to detect and analyze ADRs to minimize clinical risks and reduce healthcare costs associated with drug prescriptions. Efficient ADR-reporting processes are necessary for the timely detection of ADRs, which is crucial for preventing patient harm.

This paper aimed at comparing the effectiveness of different machine learning algorithms in detecting adverse drug reactions from social media. To ensure accurate results, various data preprocessing techniques were applied to clean and standardize the data, followed by feature extraction to vectorize it. However, the presence of imbalanced datasets can lead to overfitting. To tackle this challenge and improve the models' reliability, we employed both oversampling using SMOTE technique and undersampling to balance the datasets. This approach helps enhance the overall performance and robustness of the models in handling ADR classification. The results exhibit a consistent superiority of the Naive Bayes classifier over other classifiers, demonstrating remarkable performance across diverse datasets. Notably, it achieved impressive accuracy rates of 94.29%, 78.76%, and 64.93% on the CADEC, ADE, and PubMed datasets, respectively. The results emphasize the significance of selecting appropriate algorithms and highlight the potential of machine learning techniques for tackling complex tasks.

In the future, our goal is to develop a more comprehensive approach to ADR detection that involves not only identifying sentences containing ADRs but also extracting the specific mentions of ADRs from social media posts. This would require more advanced natural language processing techniques and could lead to more accurate and detailed insights into ADR patterns in social media data. Particularly, we plan to use large language models for extracting semantic features.

REFERENCES

[1] A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, and G. Gonzalez, "Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features," the American Medical Informatics Association, March 2015.

[2] R. Xu, Q.Q. Wang, Large-scale combining signals from both biomedical literatures and the FDA Adverse Event Reporting System (FAERS) to improve post-marketing drug safety signal detection., BMC Bioinf, Jan 2014.

[3] J. Sultana, P. Cutroneo, and G. Trifirò, "Clinical and economic burden of adverse drug reactions.," pharmacology and pharmacotherapeuticsc, 2013.

[4] M. Kuhn, M. Campillos, I. Letunic, L. Juhl Jensen and P. Bork, "A side effect resource to capture phenotypic effects of drugs," Molecular Systems Biology, 2010.

[5] B.W. Chee, R. Berlin, B. Schatz, "Identifying potential adverse effects using the web: A new approach to medical hypothesis generation," Journal of Biomedical Informatics, 2011.

[6] J. Bian, U. Topaloglu, Fan Yu, "Towards large-scale twitter mining for drug-related adverse events,", Workshop on Smart Health and Wellbeing, 2012.

[7] M. Yang, X. Wang, M. Kiang "Identification of Consumer Adverse Drug Reaction Messages on Social Media,", PACIS 2013 PROCEEDINGS, 2013.

[8] Gurulingappa and J Fluck. "Identification of adverse drug event assertive sentences in medical case reports,", In 1st international workshop on knowledge discovery and health care management (KD-HCM), 2011.

[9] H. Gurulingappa, A. Mateen-Rajpu and L. Toldo"Extraction of potential adverse drug events from medical case reports.,", Journal of Biomedical Semantics, 2012.

[10] A. Sarker, G. Gonzalez"Portable automatic text classification for adverse drug reaction detection via multi-corpus training,", Journal of Biomedical Informatics, 2015.

[11] I. Alimova and E. Tutubalina"Automated Detection of Adverse Drug Reactions from Social Media Posts with Machine Learning,", Springer International Publishing, 2018.

[12] A. Sinha, M.Nazma B.J. Naskar, M. Pandey and S. Swarup Rautaray"Text Classification Using Machine Learning Techniques: Comparative Analysis", IEEE, OITS International Conference on Information Technology (OCIT), 2022.

[13] Y. Zhanga, S. Cuic, H. Gao"Adverse drug reaction detection on social media with deep linguistic features,", Journal of Biomedical Informatics, 2020.

[14] Swati Dongre and Jitendra Agrawal "Deep-Learning-Based Drug Recommendation and ADR Detection Healthcare Model on Social Media,", IEEE Transactions on Computational Social Systems, 2023.

[15] N. Alvaro, Y. Miyao, and N. Collier"TwiMed: Twitter and PubMed Comparable Corpus of Drugs, Diseases, Symptoms, and Their Relations,", JMIR Public Health Surveillance, 2017.

[16] S. Karimi, A. Metke-Jimenez, M. Kemp, C. Wang "Cadec: A corpus of adverse drug event annotations,", Journal of Biomedical Informatics, 2015.

[17] Gurulingappa H, Rajput AM, Roberts A, Fluck J, Hofmann-Apitius M, Toldo L. "Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports,", Biomed Inform, 2012.

[18] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: Synthetic Minority Over-sampling Technique,", Journal of Artificial Intelligence Research, 2002.