Data-driven solution to identify sentiments from online drug reviews.

HAQUE, R., LASKAR, S.H., KHUSHBU, K.G., HASAN, M.J. and UDDIN, J.

2023

Article

Data-Driven Solution to Identify Sentiments from Online Drug Reviews

Rezaul Haque [1] , Saddam Hossain Laskar [1] , Katura Gania Khushbu [1] , Md Junayed Hasan [2] and Jia Uddin [3,*]

1 Department of Computer Science and Engineering, East West University, Dhaka 1212, Bangladesh
2 National Subsea Centre, Robert Gordon University, Aberdeen AB10 7AQ, UK
3 Artificial Intelligence and Big Data Department, Endicott College, Woosong University, 171 Dongdaejeon-ro (155-3 Jayang-dong), Daejeon 300718, Republic of Korea
* Correspondence: jia.uddin@wsu.ac.kr

Abstract: With the proliferation of the internet, social networking sites have become a primary source of user-generated content, including vast amounts of information about medications, diagnoses, treatments, and disorders. Comments on previously used medicines, contained within these data, can be leveraged to identify crucial adverse drug reactions, and machine learning (ML) approaches such as sentiment analysis (SA) can be employed to derive valuable insights. However, given the sheer volume of comments, it is often impractical for consumers to manually review all of them before determining a purchase decision. Therefore, drug assessments can serve as a valuable source of medical information for both healthcare professionals and the general public, aiding in decision making and improving public monitoring systems by revealing collective experiences. Nonetheless, the unstructured and linguistic nature of the comments poses a significant challenge for effective categorization, with previous studies having utilized machine and deep learning (DL) algorithms to address this challenge. Despite both approaches showing promising results, DL classifiers outperformed ML classifiers in previous studies. Therefore, the objective of our study was to improve upon earlier research by applying SA to medication reviews and training five ML algorithms on two distinct feature extractions and four DL classifiers on two different word-embedding approaches to obtain higher categorization scores. Our findings indicated that the random forest trained on the count vectorizer outperformed all other ML algorithms, achieving an accuracy and F1 score of 96.65% and 96.42%, respectively. Furthermore, the bidirectional LSTM (Bi-LSTM) model trained on GloVe embedding resulted in an even better accuracy and F1 score, reaching 97.40% and 97.42%, respectively. Hence, by utilizing appropriate natural language processing and ML algorithms, we were able to achieve superior results compared to earlier studies.

Keywords: deep learning; word embedding; Bi-LSTM; GloVe; drug sentiment analysis; drug discovery

1. Introduction

The emergence of Web 2.0 made it possible for the internet to become a more participatory platform for its users, there now being large volumes of user-generated content on social networking and online shopping websites [1]. The data growth of these platforms has been phenomenal, and the significant influence that it has on users and their families is being investigated by an increasing number of sectors in order to acquire knowledge into their user communities and, as a corollary, to drive change. Massive amounts of online data created from user comments are assessed autonomously in the pharmaceutical and healthcare industries in order to obtain helpful insights on the efficacy and side effects of medicines [2,3]. These insights are obtained from the evaluation of large volumes of online data. Online evaluations or observations regarding prescription or nonprescription pharmaceuticals, dietary supplements, and other wellness products, otherwise known as online drug reviews, have become increasingly prevalent on the World Wide Web. These

reviews, which are frequently submitted by individuals who have had firsthand experience with the product in question, provide a personal perspective regarding the product's efficacy, potential adverse reactions, and overall satisfaction. Online drug reviews can be found on websites dedicated to health and wellness, as well as on forums, social media, and e-commerce websites that sell health products [4]. People use online drug reviews as one of several sources of information when considering a health product, alongside always consulting with a healthcare provider before using any new medication or supplement. They can provide valuable information to individuals who are considering using a particular product, but it is important to keep in mind that these reviews are not necessarily representative of everyone's experience with the product, and may not always be accurate or impartial.

A sentiment analysis (SA) of drug reviews can provide valuable insights into the experiences and opinions of individuals who have used specific medications, supplements, or other health products [5]. Using ML algorithms, SA can automatically classify drug reviews based on the sentiment expressed in the text. People post their thoughts about the effectiveness or side effects of drugs on different online forums, and due to the immense growth of reviews, it has become a challenging task to extract accurate sentiments using SA. By analyzing the underlying emotional tone or attitude expressed in these reviews, SA can provide a broad and diverse perspective on the effectiveness, side effects, and overall satisfaction with a particular product [6,7].

There are several reasons why SA is an important tool for drug reviews. Firstly, it provides improved product understanding. By analyzing the sentiment expressed in online drug reviews, healthcare providers and manufacturers can gain a more comprehensive understanding of the strengths and weaknesses of their products [8]. This information can inform product development and improvement efforts, and help to ensure that products meet the needs and expectations of patients and consumers. Secondly, SA provides valuable customer feedback. Online drug reviews are a valuable source of feedback from customers, and SA can help to identify patterns and trends in this feedback. This information can inform product development, marketing, and customer support efforts, and help to improve the overall customer experience [9,10]. Thirdly, SA can provide market intelligence. By analyzing the opinions and attitudes of customers in a particular market, SA can provide valuable insights into the market landscape and inform competitive analysis. Industry manufacturers use the information to enhance the performance of their pharmacovigilance systems by identifying issues faster, comparing the online reputation of brands, posting marketing drug surveillance, and providing safe drugs without any side effects. This can help healthcare providers and manufacturers to determine informed decisions about product positioning, marketing, and regulatory efforts. Finally, SA serves as an early warning system. By identifying negative opinions and experiences with a product in a timely manner, SA can help healthcare providers and manufacturers to address potential problems before they escalate. The detection of consumers suffering from adversarial drug reactions can result in many lives being saved, which can be achieved by mining online user reviews towards a particular brand or drug. The early indication of adverse drug reactions can be helpful in maintaining customer satisfaction and protecting the reputation of the product and the company.

While the SA of online drug reviews has the potential to provide valuable insights into customer experiences, it also presents a number of challenges. One of the main challenges is the presence of noisy data. Online drug reviews can contain a significant amount of irrelevant information, such as typos, grammatical errors, and nonstandard language that can make it difficult for SA algorithms to accurately identify the sentiment expressed [11,12]. Another challenge is the subjectivity of the sentiment. The subjectivity of opinions and emotions expressed in different ways, influenced by factors such as the reviewer's personality, cultural background, and current mood, can make it difficult for SA algorithms to accurately categorize the sentiment expressed in a review. In addition, drug reviews can be highly personal and sensitive, as they often contain information

about an individual's health and well-being [13]. SA algorithms must be able to handle sensitive information and protect the privacy of individuals. Finally, the fast-paced nature of online drug reviews can also pose a challenge. New reviews are added constantly, and SA algorithms must be able to quickly and accurately analyze large amounts of new data. This requires the use of scalable and efficient algorithms that can handle large amounts of data in real time.

As a result, over the course of many years, a significant amount of research on the SA of drug reviews has been carried out to gain an understanding of a patient's level of satisfaction regarding factors such as contentment, surroundings, accessibility, the cordiality of the staff, and the effectiveness of the proceedings. In [14], the authors present a drug recommendation system that sorts sentiments into binary classes using ML models trained on different feature extractors. The authors of [15] performed a binary-class (positive and negative) and multiclass (positive, neutral, and negative) sentiment classification using a number of different ML techniques to assess the amount of efficiency possessed by a certain medication. In addition, a fuzzy-rough feature selection technique was utilized by authors of [16] to train a ML classifier to predict multiclass sentiments on drug reviews. Most of the research works were conducted to perform a binary or multiclass classification on a dataset collected from the UCI ML repository of "drugs.com" [5]. Previous research works commonly used the dataset to train different feature extractors, such as the term frequency–inverse document frequency (TF–IDF), Word2Vec, and BoW, on traditional ML algorithms to sort sentiments into positive–negative or negative–neutral–positive groups. However, their models suffered from a low accuracy score due to the imbalanced distribution of classed, high-dimensionality feature vector, and a lack of preprocessing techniques. The authors of [17] solved the low classification score problem, achieving a 93% accuracy score on the multiclass group through training an artificial neural network (ANN) model on a count vectorizer (CV) feature extractor. The study suggests that the adaptation of DL classifiers can result in a significant classification score, along with performance for the task of conducting a drug SA. Similar works could be seen performed by the authors of [18], where several combinations of DL models, namely, CNN, LSTM, Bi-LSTM, and BERT, were trained to sort drug reviews into three classes.

Though much research has been conducted on this in previous years, there are some areas where further research is needed to advance the field of drug SAs and to improve our understanding of patient experiences with medications. These include the development of more robust algorithms, integration with other sources of information, privacy protection, real-time analysis, and the interpretability and explainability of results. One of the main gaps is the development of robust and accurate SA algorithms. Despite the rapid development of ML algorithms, there is still room for improvement in the accuracy of SA algorithms when it comes to online drug reviews. The utilization of DL algorithms in prior research endeavors centered on the analysis of sentiment in regard to drugs has become quite prevalent. Despite this widespread use, these DL classifiers have exhibited a number of limitations, such as the requirement for a copious amount of data, the necessitation of human involvement, a substantial computational expenditure, and a high degree of sensitivity to parameters, all of which make the process challenging to debug. Contrarily, ML models, though shown not to be as accurate as DL classifiers in prior research, necessitate fewer computational resources and a reduced amount of human involvement [19]. The efficacy of ML algorithms in drug SAs is critically dependent on the text preprocessing and feature extraction methods employed [20]. As a result, it is imperative to properly select the appropriate text-cleansing technique and feature extractor, among the various alternatives available, to enhance the performance of ML algorithms in the task of conducting drug SAs. Further research is imperative to augment the performance of ML algorithms in drug SAs.

As a result, the current study was designed to develop both ML and DL algorithms that could better handle noisy data and subjectivity in online drug reviews to provide more accurate results. In our pursuit to achieve the stated objective, we provided several noteworthy contributions.

- Our work contributes to the advancement of the field of drug SAs by providing a comprehensive comparative analysis of ML and DL algorithms, and a valuable resource in the form of a large corpus of labeled drug reviews.
- Our web application provides valuable insights into consumers' experiences with drugs, which pharmaceutical companies can use to improve their products and services.

The paper is organized into the following sections: Section 2 describes related works; Section 3 contains the methodology of the research, which consists of data collection and labeling, utilized text preprocessing, feature extraction techniques, and training parameters of ML models, as well as opted evaluation metrics; Sections 4 and 5 contain the result analysis and discussion, respectively. In the final section, Section 6, we conclude our research work, providing the limitations and potential future works.

2. Related Works

There has been significant effort put into utilizing ML/DL algorithms to discern the feelings of user evaluations, which coincides with the dramatic increase in improvements of AI. The restaurant, e-commerce, and other industries frequently utilize SA to understand the opinions of consumers to grow their business. In spite of the widespread use of SA across a wide variety of application areas, the pharmaceutical domain has received a significantly smaller amount of attention. On the other hand, many developments have been recorded in the more recent literature. This is due to the relevance of mining medication reviews, which could contribute to a variety of different healthcare stakeholders.

Many research studies implement drug SAs using different preprocessing and feature extraction techniques. Unfortunately, there is a lack of research works on drug SAs with great accuracy scores due to the use of inappropriate preprocessing and feature extraction techniques, as well as unsuited training parameters. Due to the lack of ground-truth datasets, most of them used datasets from drugs.com and trained ML/DL algorithms to perform either binary (positive and negative) or multiclass (positive, neutral, and negative) classifications. The study in [14] presents a drug recommendation system based on ML algorithms such as LR, perceptron, ridge classifier, multinomial naïve Bayes, SGD, and SVM, where binary classification is performed to identify sentiments in positive or negative circumstances. To train the ML models, they used four feature extractors and their LR model, which resulted in the highest accuracy score of 91%. On a similar dataset, the authors of [16] proposed a fuzzy-rough feature-selection-based ML model to classify sentiments into three classes. They used BOW and TF–IDF to train naïve Bayes, random forest, decision tree, and riper models, where a random forest method with TF–IDF obtained the highest accuracy of 67%. In [21], the authors proposed a linguistic approach for conducting a drug SA on a multiclass dataset, which was collected from WebMD. Their approach outperformed two types of SVM models with an accuracy of 69%, exceeding the score of 7%. The authors of [22] investigated the effect of SA features in detecting adversarial drug reactions from online posts. They created a dataset from Twitter and DailyStrength and performed a binary classification to achieve an 80% accuracy score.

In recent years, DL algorithms have emerged as the most popular technique to use with drug SAs. In [18], the authors trained both ML and DL models with different feature extractors, such as TF–IDF, CV, and Word2Vec, to classify drug sentiments into a multiclass classification. On the testing data, their ANN model obtained the highest accuracy score of 93.85% with a CV. The authors of [18] conducted a similar work, where a comparison of several DL classifiers' performance on a multiclass drug SA was presented. They trained a CNN, LSTM, Bi-LSTM, BERT, and a combination of these models on Word2Vec word embedding, achieving the highest F1 score of 0.90 with a combined model of BERT and LSTM. These two papers obtained significant classifications for the drug SA task using DL algorithms, but their ML models resulted in poor performance on testing data.

It is abundantly obvious from the published works that the ML models of earlier works did not have a substantial accuracy score when using ML algorithms. Therefore, it is essential to devise appropriate methods for enhancing the performance of these models

in order to meet the demands of the field. Additionally, we believe that with the proper selection of a preprocessing method, feature extraction, and ML models, the results of the previous results can be improved. According to our best knowledge, there is no research on multiclass SAs of drug reviews obtained from drugs.com that had a great accuracy score using ML algorithms.

3. Methodology

Figure 1 shows the proposed methodology of the current study. Firstly, we collected a substantial corpus of 215,063 drug reviews from the drugs.com website and categorically labeled them into three classes, namely, positive, negative, and neutral. Subsequently, we carried out a comprehensive text preprocessing procedure, leveraging the capabilities of Python's Natural Language Toolkit (NLTK) package to minimize the presence of unwanted noise in the text and to improve its overall quality. Furthermore, we trained five different ML algorithms, namely, random forest (RF), support vector machine (SVM), passive aggressive (PAG), logistic regression (LR), and stochastic gradient descent (SGD), on two feature extractors, TF–IDF and CV, to perform a SA on the drug reviews. Additionally, we trained four DL algorithms, LSTM, Bi-LSTM, GRU, and Bidirectional GRU (Bi-GRU), on two word-embedding techniques, word2sequence and GloVe, to perform a SA on the drug reviews. We also conducted a comparative analysis of the performance of the classifiers, based on accuracy, error analysis, and model performance, to determine the optimal classifier for the SA on the drug reviews. Finally, we built a real-world web application, utilizing the Flask framework, based on the classifier with the highest performance, which could sort drug reviews into the three classes. The real-world web application for the drug SA has the potential to greatly impact the healthcare industry and provide valuable insights into consumers' experiences with drugs.
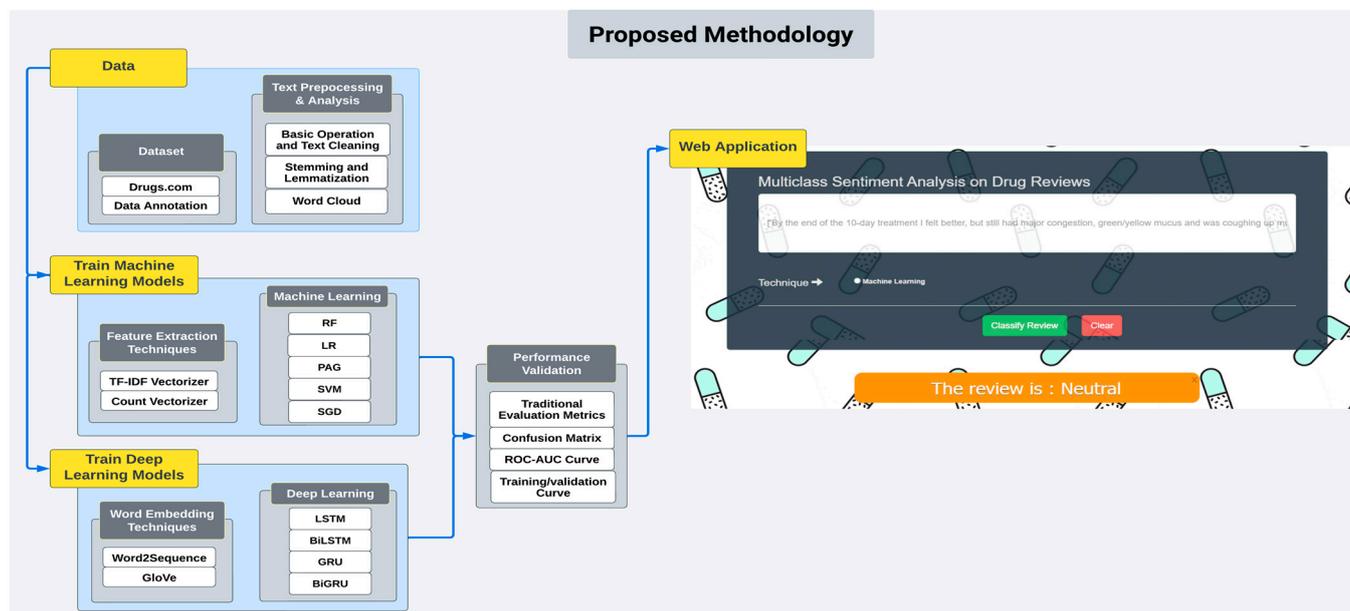

Figure 1. Proposed methodology.

3.1. Data Collection

The experimental dataset used for the drug SA was referred to as the drug review dataset, and was obtained from the UCI repository. It consisted of user evaluations of different medications, related ailments, and a range of star ratings that indicated the level of user satisfaction. The dataset also included the drug's name, patient's condition, useful counts, and the number of people who found the review helpful. Each drug review was rated on a scale of 0 to 9, with 0 representing the least satisfied patients and 9 representing

the most satisfied patients. Based on the review's rating, the dataset was categorized into three classes, namely, negative (rating less than 4), neutral (rating greater than 4 and less than 7), and positive (rating greater than 7). The distribution of the labels for each class can be seen in Figure 2. The final dataset contained 215,063 drug reviews, where positive, negative, and neutral classes contained 105,433, 100,071, and 9559 reviews, respectively.
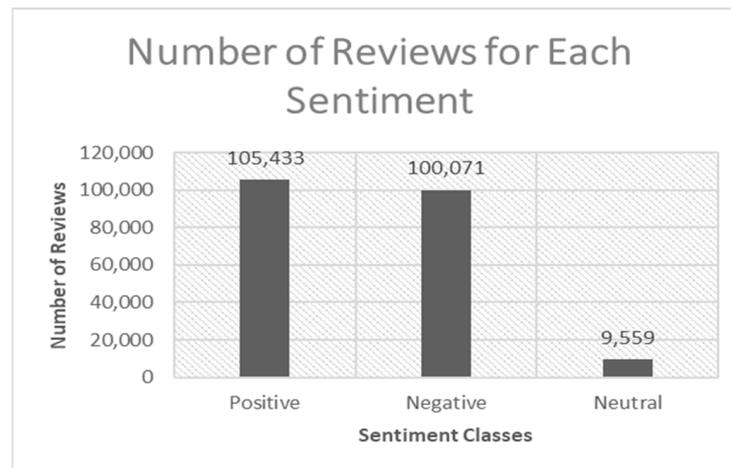

Figure 2. Number of reviews for each class.

Table 1 shows the statistical analysis of the drug review dataset used in the study. The table shows the number of reviews, total words, and unique words in each of the three classes: positive, negative, and neutral. The positive class had the highest number of reviews with 105,433, followed by the negative class with 100,071, and, finally, the neutral class, with 9559 reviews. Additionally, the positive class had the highest number of total words, with 4,145,737, followed by the negative class with 4,319,347, and the lowest number of total words was recorded in the neutral class, with 230,376. Finally, the number of unique words for each class was also shown, where the positive class had the highest number of unique words, with 30,752, followed by the negative class with 31,418, and the neutral class had the lowest number of unique words, with 8561.

Table 1. Statistical analysis of the drug review dataset used in the study.

Class	Total Reviews	Total Words	Unique Words
Positive	105,433	4,145,737	30,752
Negative	100,071	4,319,347	31,418
Neutral	9559	230,376	8561

3.2. Text Preprocessing

In SA, text preprocessing plays a crucial role in improving the accuracy of the models applied to the text. Text obtained from the internet often contains a high amount of noise, including advertisements, HTML elements, scripts, punctuation, and white space. Data preprocessing helps eliminate incomplete, noisy, and inconsistent data. By removing all these elements, the amount of noise in the text was reduced, leading to the better performance and precision of the classification models applied. In this study, special data-cleaning steps were performed, such as stopping words from being removed and digit removal, as well as text normalization, where contractions were expanded and lemmatization and spelling corrections were applied to reduce the dimensionality of the data. However, care must be taken during the noise removal and text normalization processes, as they can result in the loss of a small number of rows from the dataset, potentially reducing the accuracy. Figure 3 shows the word cloud of the most frequently occurring tokens in each class after

all the preprocessing steps were completed. Figure 4 shows the most frequent words in the experimental dataset.
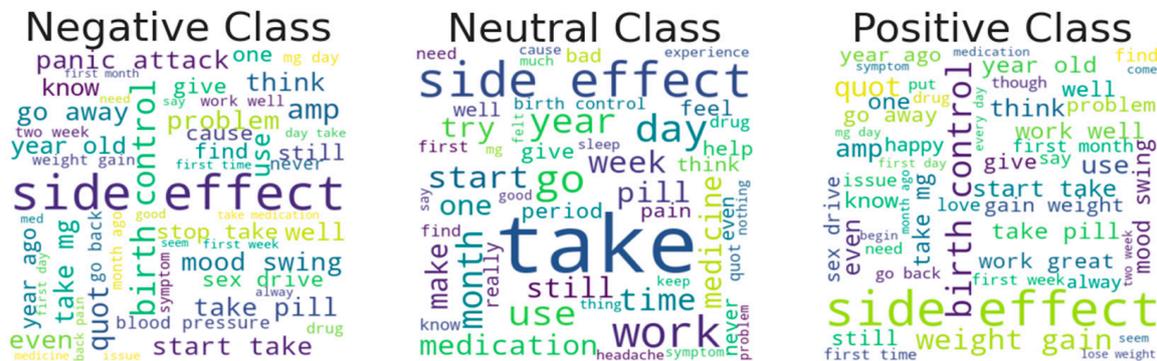

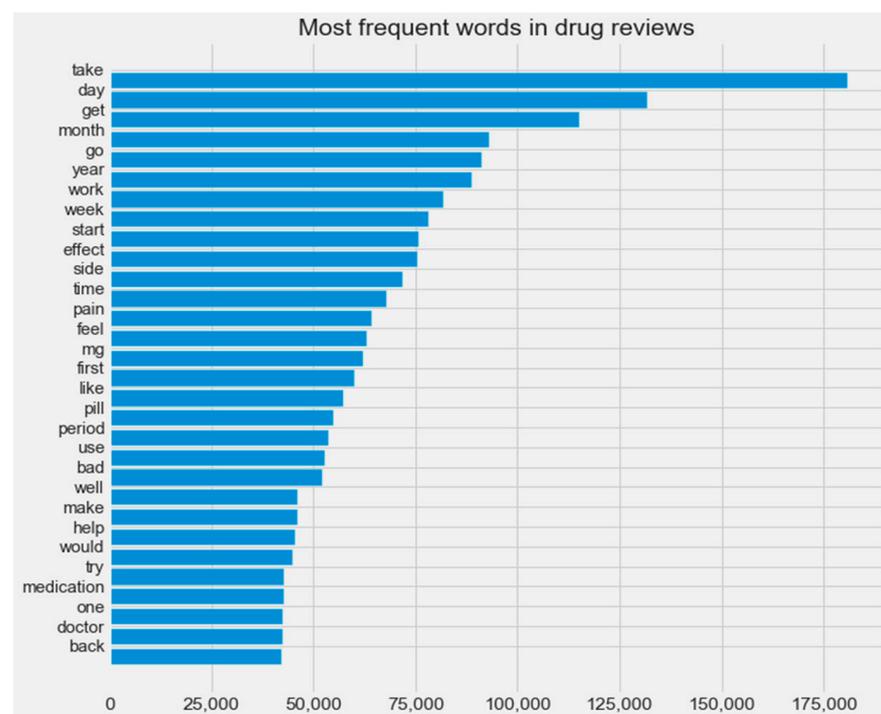**Figure 3.** Word cloud of most frequently occurring terms for each class of reviews.

Figure 4. Most frequent words in the experimental dataset.

3.3. Feature Extraction

We utilized feature extraction techniques, such as TF–IDF and CV, as well as word-embedding techniques, such as word2sequence and GloVe embedding, to convert raw text into numerical features for the ML and DL algorithms. TF–IDF assigns weight to words based on their frequency of occurrence in a document and the entire dataset, while a CV counts the number of times each word appears in a document [23,24].

3.3.1. TF–IDF Vectorizer

TF–IDF is a widely-used method for determining the importance of words in a document. Term frequency is calculated by dividing the number of times a word appears in the text by the total number of words in the document [25]. IDF, which stands for "inverse document frequency", calculates the importance of a phrase. It is calculated using the formula IDF (t) = log(N/DF), where N is the total number of documents and DF is the number of documents that contain the phrase t. The transformation of information from its

narrative structure into a vector space can be performed more effectively using TF–IDF. This allows for the location of phrases inside a text that are essential and carry a lot of weight. The following Equations (1)–(3) were the formulae for TF–IDF, where *i* refers to the word and *j* refers to the document. The supervised ML algorithms used N-Gram to derive the text's characteristics. N-Gram represents the n tokens sequentially taken from the provided text, where n can have values of 1, 2, 3, 4, etc. A unigram, bigram, trigram, etc., correspond to n values of 1, 2, 3, etc., respectively. In this study, the sklearn feature extraction library was used to apply TF–IDF on cleaned reviews. The N-Gram range was set to unigram and bigram, with a maximum feature limit of 12,000.

$$TF(i,j) = \frac{Term\ i\ frequency\ in\ dcument\ j}{Total\ words\ in\ document} \tag{1}$$

$$IDF(i) = \log_2\left(\frac{Total\ documents}{documents\ with\ term\ i}\right) \tag{2}$$

$$TFIDF(i) = TF(i,j) \times IDF(i) \tag{3}$$

3.3.2. CV

High-frequency terms were chosen for the vocabulary via the CV method, which then built a sparse representation of the texts across the vocabulary. The documents were transformed into a matrix that contained token counts by using the CV. To begin, the documents were segmented, and then a dense matrix was constructed based on the frequency with which each token appeared in the texts. In order to construct the matrix, we first filtered out all of the document collection's stop words. The next step was to clean up the vocabulary by removing any terms that were found in less than four different texts. Through this method, any phrases that were used insufficiently were eliminated. In this particular research study, the CV was applied to cleaned reviews utilizing a unigram and bigram N-Gram range with a default maximum feature parameter set.

3.3.3. Word Embedding

The innovation of word embedding has highly improved the performance of DL classifiers for the task of text classification [26]. Word embedding is an improved version of the bag-of-words method that maps tokens of textual data into a dense vector representation of words. It can extract relative and semantic information from the text of a collection of statistical language modeling techniques and provide a continuous vector space for DL classifiers to work on [27]. DL classifiers take the data into input, hidden, and output layers to extract meaningful features from the text. The embedding layer is fed onto the input layer of the model to learn from the vectorized words. For the task of text classification, there are different types of word embedding techniques, and among them, word2sequence and GloVe were utilized in this experiment, both of which could capture the semantic meaning and relationships between words.

The first step in the process of the SA was to convert the textual data into numerical inputs that could be processed with the ML/DL algorithms. To achieve this, two feature extraction techniques were implemented: the TF–IDF vectorizer and CV. Additionally, two word-embedding techniques were used to train the DL classifiers: word2sequence and GloVe embedding.

The word2sequence embedding technique was utilized to vectorize the text corpus by converting each text into a sequence of integers. This was performed using word2sequence's tokenizer and padding functions, where each integer represents the index of a token in a dictionary. The dataset was tokenized with a vocabulary size of 43,000 and an embedding dimension of 200, since a higher dimension could result in computational difficulties for the DL classifiers. The length of the text sequence was maintained equal by adding padding zeros at the end, as the lengths of the comments varied.

We also employed pretrained GloVe embedding, an unsupervised learning approach that could generate word representations for global vectors. It is a pretrained technique that is highly efficient due to being trained on a large corpus of text. The technique is simply a log-bilinear approach with a scaled least-squares purpose that was developed on a 6-billion-token corpora. The corpus was built with Wikipedia2014 and Gigaword5, with a vocabulary of the top 400,000 most frequently occurring terms and a context window size of 10. In this experiment, we used word vectors with dimensions of 300, vocabulary sizes of 43,000 tokens, and padding sizes of 100 to initialize the GloVe embedding.

3.4. Training Baseline Models

In our study, we selected a baseline model as a crucial step to contextualize the results of our trained models. The baseline model served as a comparison point to evaluate the performance of the classifiers. To determine the baseline model, we selected the highest performing ML classifier, which was the RF classifier trained using the CV feature extraction method. Before training the ML model, we split the dataset into training and testing sets, with 80% of the data being used for training and 20% for testing. The training set comprised of labeled drug reviews, while the testing set consisted of unlabeled reviews. This helped us to train our model using the labeled drug reviews and evaluate its performance on the testing set, which consisted of a smaller portion of data. Descriptions of the opted for ML models are described in what follows.

3.4.1. RF

The classification technique that consists of many decision trees is referred to as the RF classifier. The algorithm uses randomness to create each tree and develop independent forests. These forests are then used to determine accurate predictions [28]. Two main parameters must be configured to implement the random forest classifier: the number of estimators and the criteria used. According to the findings of a number of research works [29], it is possible to attain good outcomes by sticking with the system's default settings. However, the enormous number of trees produces a consistent outcome of varying relevance, as stated in [30]. In addition, [31] claimed that utilizing a greater number of trees than what is required may be superfluous, but that this does not decrease the model's accuracy. Both the n estimator parameter (which was set to 100), as well as the criteria parameter (which was set to entropy), went through a series of tests and evaluations in this study in order to identify the best possible RF model for performing the classification.

3.4.2. SVM

SVM is an effective approach that can serve both the objectives of regression and classification, and involves drawing a hyperplane to demarcate separate categories. SVMs function admirably even when the number of observations is far higher than the dimension size, but one disadvantage is that they do not work very well with large datasets. The authors of [32] suggested that the RBF kernel of an SVM classifier yields good results. When implementing this method, two criteria need to be specified: the cost parameter (C) and the kernel width parameter. The C parameter adjusts the level of stiffness for nonseparable training data, while the kernel width parameter affects the refinement of the class-dividing hyperplane. Increasing the value of C may lead to overfitting, while increasing the value of the kernel width parameter may affect the accuracy of the classification. When we trained the SVM model, we set the kernel type to be RBF, and provided the cost parameter a value of 1.

3.4.3. PAG

PAG algorithms are a class of ML algorithms that are frequently utilized in software designed for working with large amounts of data [33]. The algorithm is frequently used for numerous kinds of large-scale and online learning. In online ML methods, the input data arrives in the order that it was requested, and the ML model is updated in the same

order. This is in contrast to standard batch learning, which uses the full training dataset simultaneously [34]. The PAG model was trained with the default parameters, and the maximum number of iterations was set to 200.

3.4.4. LR

A logistic function is used to describe a binary dependent variable in the most fundamental version of the statistical model known as the LR [35]. It is a commonly used approach for classifying data, and is a member of the generalized linear models subclass. It is an ensemble learning algorithm that predicts the probabilities that describe the results of an experiment. We trained the LR model with the cost parameter set to 1, the maximum iteration set to 200, the tolerance set to 0.001, and the liblinear solver settings.

3.4.5. SGD

Commonly used in neural networks, SVMs, and LRs, the following algorithm is well adapted to permit the discriminative training of linear models under convex loss functions [36]. SGD is a common ML method for model optimization, which is an advanced version of gradient descent. It is a stochastic estimation of the gradient descent optimization method that uses an interpolation of gradients by subsampling the entire training set of data. The approach is widely used due to the fact that it has a high efficiency and is simple to construct for datasets that contain redundant observations [37]. In this work, an LR was utilized as a loss function in the modeling, and the maximum number of iterations that were used was 200.

3.5. Training DL Models

We divided our dataset into 75% training, 5% validation, and 20% testing sets, where the training and validation sets were used to train the LSTM, Bi-LSTM, GRU, and Bi-GRU models. The testing set would be used to determine predictions on the model after training in order to compare the effectiveness of the classifiers. However, all of these classifiers have their own advantages and disadvantages. The DNN network is highly sensitive to parameters, and so, it was necessary to find the proper parameters to train them.

There are numerous RNN-based model variations that work well with sequential input data, including audio, music, text, name entity recognition, etc. However, due to the vanishing and exploiting gradient problem, it often does not perform well in long-term dependence [38]. Since the network's half that is closest to the output is updated, the other half that is farther away from the output is improperly updated. That is why LSTM, a particular kind of RNN that can learn long-term dependencies, was developed. They can develop a future model based on past and present data by learning from past data, which is accomplished by incorporating a number of gates into their network design to recall past data. As a result, the input values are only traversed once (i.e., from left to right, input to output). Additionally, GRUs have been utilized recently in order to alleviate the shortcomings of standard RNNs with large texts [39]. The advantage of the GRU model is that it can be used to decide how much information should be remembered from previous steps, how much should be handed back, and how much should be received in the current synthesis steps. Through these gate controls, the GRU is able to learn long texts effectively. Consequently, the RNN networks with memory operations are obviously more appropriate for the task of text classification. However, LSTM is hard to train fast and accurately, as it takes a lot of resources [40]. In the Bi-LSTM model, the given input data were utilized twice for training (i.e., first from left to right, and then from right to left), which compensated for the shortcomings of LSTM by learning from the previous information of the current word, fully considering the semantic features between contexts and acquiring more comprehensive features and feature information. The Bi-GRU, which merges the "forget" and "input" gates into a single update gate, is a minor modification of the Bi-LSTM. Along with applying various adjustments, it blends the hidden state and cell state.

In our research, we trained these models on word2sequence and glove embedding separately. Figure 5 depicts the layers of directional and bidirectional networks corresponding to the embedding layer as the input layer, followed by the DL algorithm and the dropout layers as the hidden layer, and 3 units (positive, negative, and neutral) of output layers. With the training and validation data, the models were individually trained with batches of 128. The model was configured to train, at most, with 100 epochs. However, to avoid overfitting an early stopping method, which monitors the validation accuracy per epoch with a factor of 0.1, patience of 2, and a minimum delta of 0.0001, a minimum learning rate of 1×10^{-6} was incorporated that monitored the accuracy of the model, embedded within the training stage. As it was a multiclass problem, the loss function used for training the models was the "sparse categorical cross-entropy" and the used optimizer was "Adam", with the learning rate 1×10^{-2} and epsilon of 1×10^{-8} provided. Then, the model was evaluated on the 20% test data to determine predictions on three classes of drug reviews.
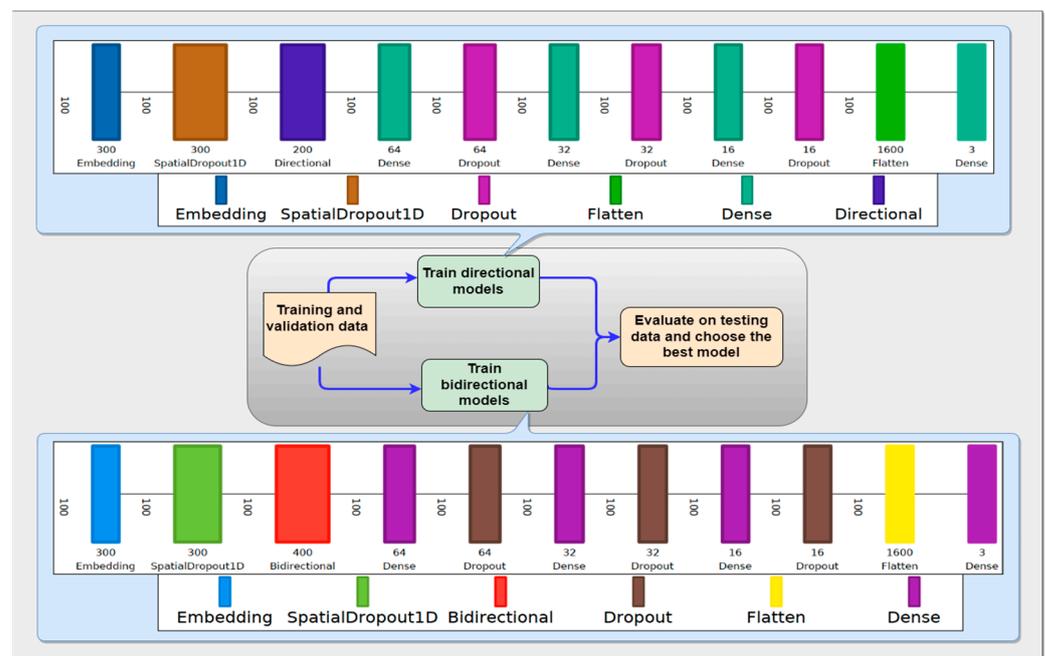

Figure 5. Neural architecture of bidirectional and directional LSTM networks.

3.6. Evaluation

The performance of our proposed models was evaluated based on their accuracy, precision, and recall score, as well as their F1 score. Equations (4)–(7) were used to calculate the traditional evaluation metrics.

$$accuracy = \frac{TP + TN}{\text{TP} + \text{FP} + \text{FN} + \text{FP}} \tag{4}$$

$$precision = \frac{TP}{\text{TP} + \text{FP}} \tag{5}$$

$$recall = \frac{TP}{\text{TP} + \text{FN}} \tag{6}$$

$$f1 - score = \frac{2 \times (precision \times recall)}{\text{precision} + \text{recall}} \tag{7}$$

In Equations (4)–(7), the number of data identified as positive among the data labeled as positive was referred to as the true positive (TP), and the number of data classed as negative among the data labeled as negative was referred to as the true negative (TN). A

false negative (FN) was the number of data that were supposed to be negative but were really labeled as positive in the dataset, while a false positive (FP) was the number of data that were supposed to be positive but were actually marked as negative in the dataset. Accuracy was defined as the number of properly predicted cases divided by the total number of predictions provided by the model; however, any approach could perform differently in relation to the number of instances that were correctly classified, as seen in Equation (4). According to Equation (5), precision was measured as the percentage of documents that were correctly identified as positive by the model. As indicated in Equation (6), the recall was the proportion of documents that were categorized as positive by using the model out of the total number of documents that really had positive tags. In addition, the F1 score was the overall average of the recall and accuracy scores, as seen in Equation (7). In addition, an error analysis was performed with the help of a confusion matrix.

4. Results Analysis

4.1. Performance of ML Model

Table 2 shows the results of the baseline models applied to the drug reviews. It was found that the RF model performed better than the other models when using either TF–IDF or the CV for the feature extraction. The models trained with the CV had a higher accuracy score compared to those trained with TF–IDF. Only the RF model showed a good classification score when trained with TF–IDF, while the other algorithms performed poorly. Furthermore, the SGD model had the worst performance for both feature extraction methods.

Table 2. Results of all the classifiers trained using TF–IDF and CV.

Feature	Models	Evaluation Metrics			
		Accuracy	Precision	Recall	F1
CV	RF	96.65%	96.60%	96.13%	96.42%
	LR	96.04%	96.94%	95.54%	96.24%
	PAG	95.57%	94.56%	95.51%	95.48%
	SVM	95.33%	94.46%	95.61%	95.52%
	SGD	95.27%	94.86%	95.41%	95.37%
TF–IDF	RF	93.29%	92.39%	94.59%	93.69%
	SVM	90.65%	89.99%	90.11%	89.77%
	PAG	90.39%	90.00%	90.90%	90.41%
	LR	89.62%	89.81%	86.99%	89.63%
	SGD	86.01%	87.02%	85.82%	86.54%

The choice of feature extractor and classifier significantly affected the performance of the sentiment analysis models, as shown in Figure 6. The CV outperformed TF–IDF, and RF achieved the highest accuracy scores among the classifiers, with an accuracy and F1 scores above 93%. The models trained on the CV showed RF and LR to have the highest accuracy scores of 96.65% and 96%, respectively. PAG and SVM also performed well, with accuracy and F1 scores above 95.3%. However, the SGD model had the lowest score, with an accuracy of 95.27%. All models achieved high classification scores on the CV features, indicating that further research should be conducted to determine the optimal combination of feature extractors and classifiers for sentiment analysis tasks.
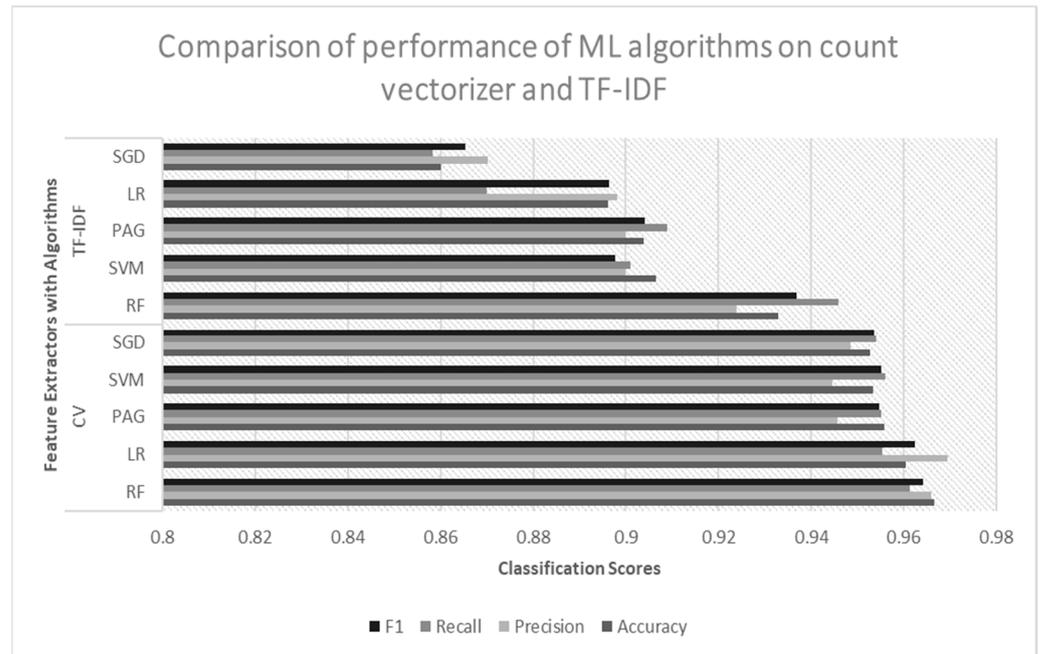
Figure 6. Results of CV and TF–IDF for each ML algorithm.

The ROC curve, also known as ROC–AUC, is a graph used in classification to display a model's performance at all classification thresholds. It helps balance sensitivity and specificity by illustrating the TPR and FPR for each threshold. The curve plots TPR on the *y*-axis and FPR on the *x*-axis, with a good model rising steeply, indicating that TPR increases faster than FPR as the probability threshold decreases. The ideal point is at the top left corner of the graph, where FPR is zero and TPR is one. The AUC–ROC takes values between 0 and 1, where 0 indicates an entirely erroneous classification and 1 indicates a completely accurate classification. Figure 7a,b show the AUC–ROC curve for RF models trained with TF–IDF and the CV, respectively. The figures suggest that the RF model trained with the CV performed better than the TF–IDF-vectorized model, achieving a mean AUC score of 98% with a great AUC score for all classes.
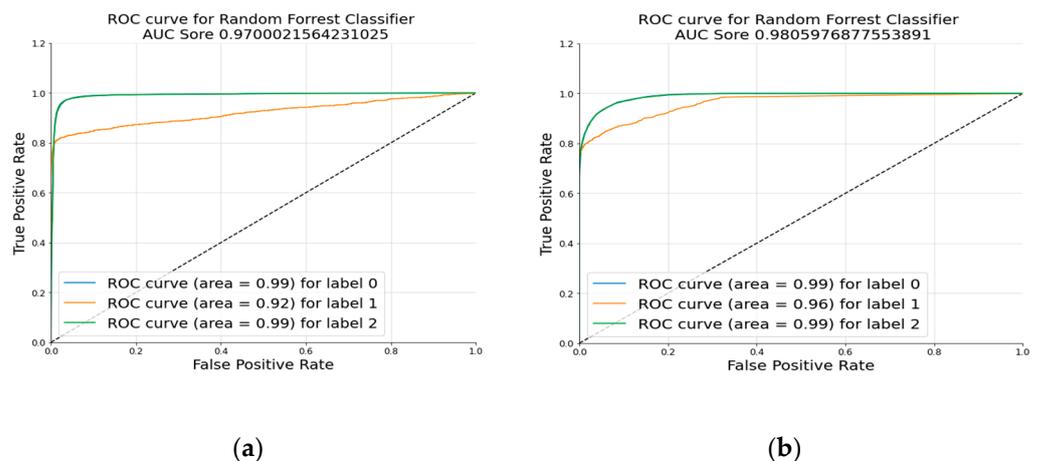

(**a**)

(**b**)

Figure 7. ROC–AUC curve of highest performing classifier trained on TF–IDF and CV: (**a**) ROC–AUC curve for RF using TF–IDF vectorizer; (**b**) ROC–AUC curve for RF using CV.

4.2. Performance of DL Classifiers

Table 3 displays the performance of the DL classifiers trained on GloVe and word2sequence embedding based on the opted for evaluation metrics. From the table, we

could see that the Bi-LSTM model trained on GloVe embedding was the best performing algorithm, outperforming our baseline and all other DL classifiers. The results of Table 3 showcased that the DL classifiers trained on GloVe embedding could achieve a significantly higher score than the models trained on word2sequence embedding. On the other hand, all the models trained on word2sequence embedding failed to beat the classification scores of the baseline model, where LSTM performed the worst among all the DL classifiers.

Table 3. Results of DL classifiers trained on GloVe and word2sequence embedding.

Feature	Models	Evaluation Metrics			
		Accuracy	Precision	Recall	F1
GloVe	Bi-LSTM	97.40%	97.01%	97.68%	97.42%
	LSTM	97.20%	97.08%	95.66%	97.34%
	GRU	95.28%	95.46%	95.51%	95.48%
	Bi-GRU	95.21%	94.82%	95.32%	95.26%
word2sequence	Bi-GRU	95.09%	95.09%	94.09%	94.59%
	Bi-LSTM	92.30%	92.30%	82.94%	87.37%
	GRU	92.21%	92.21%	82.95%	87.34%
	LSTM	91.20%	91.20%	80.18%	85.34%

Figure 8 compares the DL classifiers trained on word2sequence and GloVe embedding, showing that Bi-LSTM and Bi-GRU had the highest classification scores for GloVe and word2sequence, respectively. On the other hand, Bi-GRU (GloVe) and LSTM (word2sequence) had the overall lowest percentages in this DL classifier results. The evaluation metrics indicated that GloVe outperformed word2sequence with a 97.40% accuracy, 97.01% precision, 97.68% recall, and 97.42% F1 score. However, Bi-LSTM performed better in word2sequence embedding, ranking second in the evaluation. The Bi-GRU of word2sequence covered a 95.09% accuracy, 95.09% precision, 94.09% recall, and 94.09% F1 score, which were less than the Bi-LSTM of GloVe. Interestingly, Bi-GRU performed worst in the GloVe evaluation. LSTM ranked second, with over a 97% accuracy, precision, and F1 score, and obtained a 95.66% recall in GloVe embedding. However, LSTM in word2sequence embedding ranked last in its evaluation. GRU ranked third in both GloVe and word2sequence. However, GRU (word2sequence) failed to achieve even 90% in each evaluation segment, while GRU (GloVe) scored over 95% in each evaluation segment.

In Figure 9, the DL models trained on word2sequence embedding were compared. Bi-GRU achieved the best accuracy, precision, recall, and F1 score, with 95.09%, 94.09%, and 94.59%, respectively. Bi-LSTM and GRU had similar accuracy scores of 92%, but a low recall and F1 scores of less than 83% and 88%, respectively. LSTM had the lowest classification score with accuracy and precision scores of 91.2% and an F1 score of 85.34%. Despite Bi-GRU's high performance, it did not surpass the baseline model.

The results of the DL classifiers trained on GloVe embedding are presented in Figure 10. The Bi-LSTM and LSTM classifiers achieved the highest scores with an accuracy of 97.40% and an F1 score of 97.42%. LSTM also performed well with an accuracy score of 97.20% and an F1 score of 97.34%. GRU obtained a good score with an accuracy of 95.28% and an F1 score above 95.48%. Bi-GRU had the lowest score among the models with an accuracy score of 95.217%. Our Bi-LSTM model outperformed the baseline classifier by increasing the accuracy, recall, precision, and F1 scores by 0.75%, 0.41%, 1.55%, and 1%, respectively.
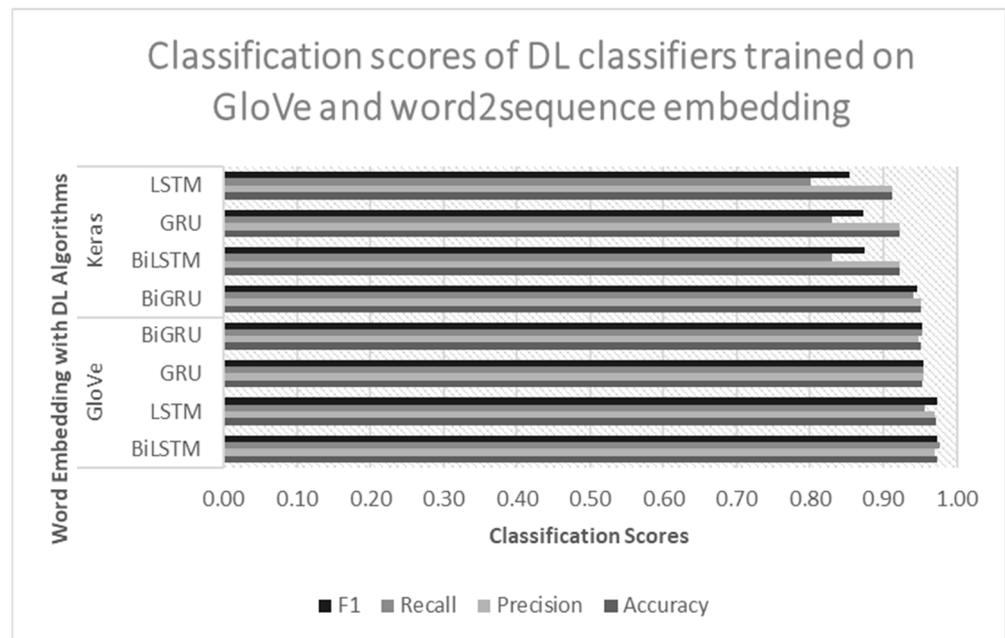
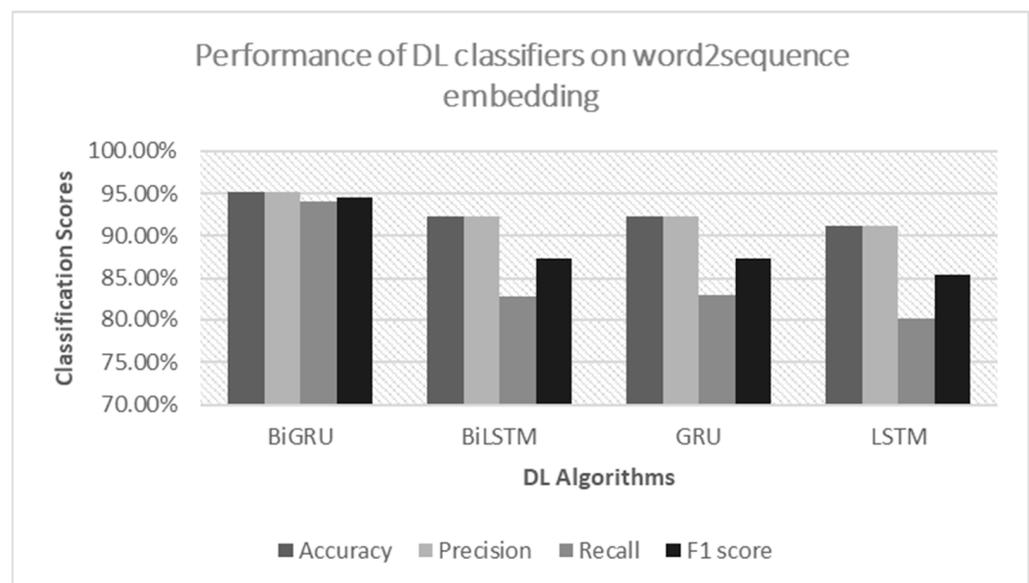**Figure 8.** Comparison of performance of DL classifiers trained on word2sequence and GloVe embedding.

Figure 9. Performance of DL classifiers on word2sequence embedding.

4.3. Error Analysis

The confusion matrix is one of the approaches that is both insightful and easy when it comes to measuring the accuracy and completeness of a ML system. Its primary application is in classification jobs, particularly those in which the results may include two or more types of classes. The technique was utilized to perform an error analysis on the highest performing model trained on each feature extraction technique. With the help of the confusion matrix, we tried to find meaningful insights about the ML/DL model results.
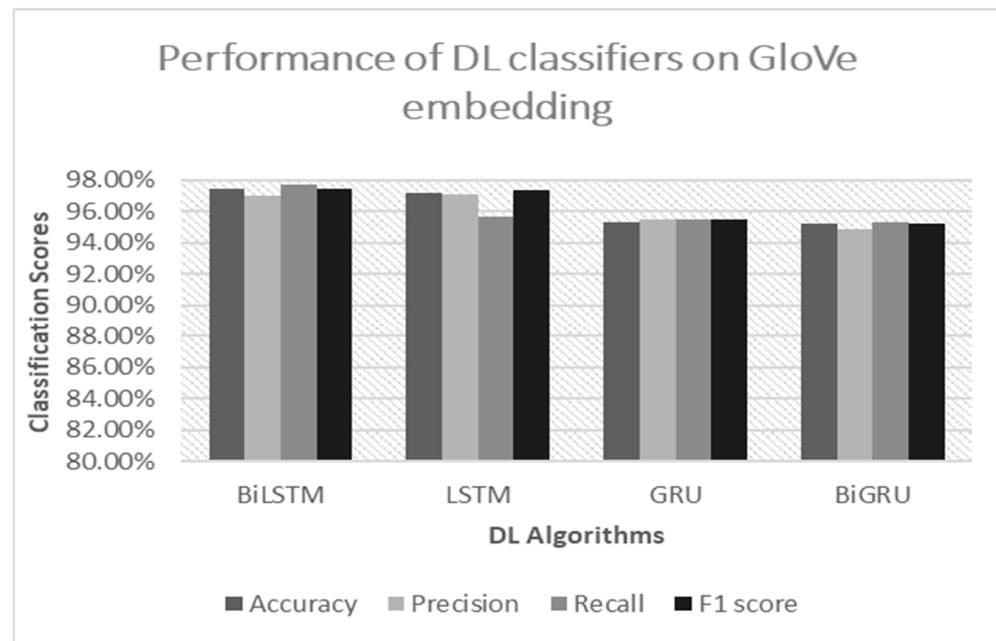
Figure 10. Performance of DL classifiers on GloVe embedding.

4.3.1. Error Analysis of ML Models

Two highest performing ML models, one trained with TF–IDF and the other with a CV, were evaluated using confusion matrices, as shown in Figure 11a,b. The RF model in Figure 11a performed better at predicting positive and negative drug reviews than neutral reviews. The count-vectorized RF model in Figure 11b performed significantly better than the TF–IDF-vectorized RF model in positive and negative classes, but both models struggled with accurate predictions for the neutral class. The count-vectorized RF model had fewer incorrect predictions and a better accuracy rate overall, with incorrect predictions for negative, neutral, and positive classes at rates of 3.2%, 25.7%, and 2.6%, respectively, on the 20% test data.
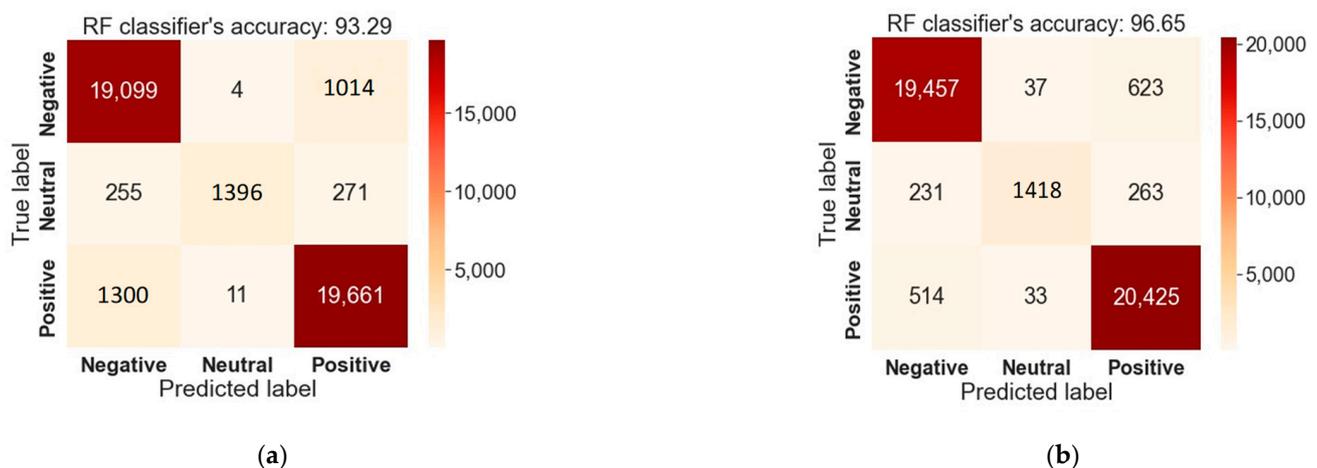

(a)

(b)

Figure 11. Confusion matrix of highest and lowest performing classifiers trained on TF–IDF vectorizer and CV: (**a**) RF trained on TF–IDF; (**b**) RF trained on CV.

4.3.2. Error Analysis of DL Models

Two highest performing models using word2sequence and GloVe embedding were evaluated, and the resulting confusion matrices are shown in Figure 12a,b. The word2sequence + Bi-GRU model correctly predicted 94.5% of the reviews, but it did not

outperform the baseline model's performance. The GloVe + Bi-LSTM model had more accurate predictions for the negative and positive classes, with only 2–2.1% incorrect predictions, but struggled with the neutral class, with 13.3% incorrect predictions. Despite this, the GloVe + Bi-LSTM model outperformed the baseline models due to its superior accuracy in predicting the negative and positive reviews.
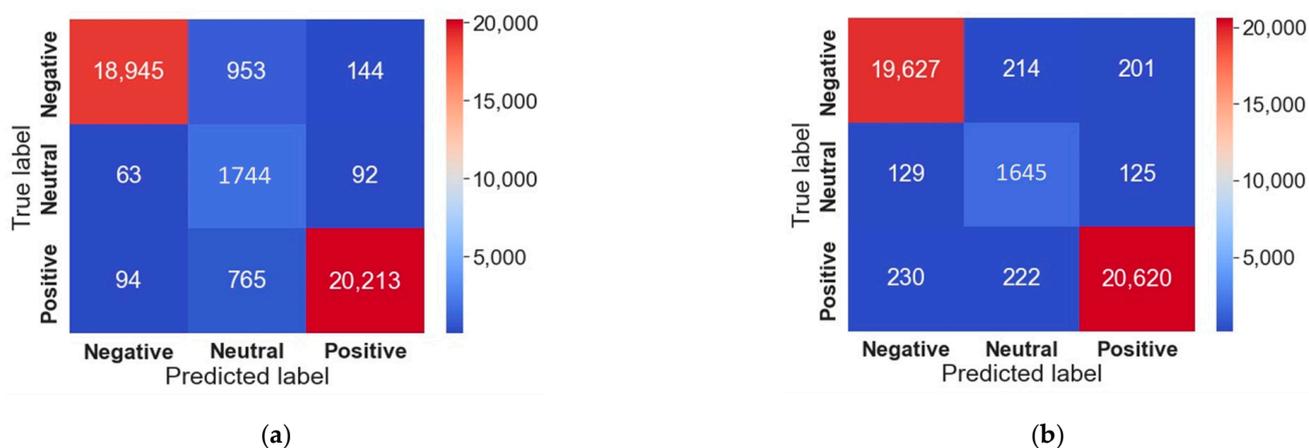

(**a**)

(**b**)

Figure 12. Confusion matrix of highest performing classifier trained on word2sequence and GloVe embedding: (**a**) Bi-GRU trained on word2sequence; (**b**) Bi-LSTM trained on GloVe.

Figure 13 illustrates the validation accuracy and validation loss for each model trained with GloVe embedding for 100 epochs. From Figure 13, we could see that the validation accuracy of Bi-LSTM and LSTM model increased with the increments in epochs, though Bi-LSTM performed slightly better with better validation accuracy scores for each epoch. Furthermore, Bi-GRU performed the worst throughout the epochs with lowest validation accuracy scores. The validation loss of each model for 100 epochs is also shown in Figure 11, showing that the loss of each model decreased until the 45th epoch, and after that, it started to overfit with an increased validation loss. The Bi-LSTM and LSTM models continued the whole process with a lower validation loss compared to the other algorithms. GRU and Bi-GRU did not perform well, as its validation loss decreased until the 45th epoch, but the loss was comparatively higher and kept going higher for the rest of the epochs.
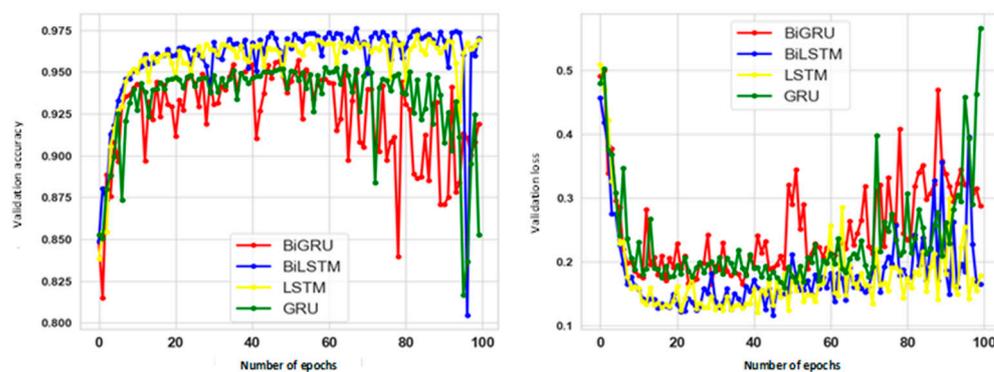

Figure 13. Comparison of validation accuracy and loss for each epoch of all the classifiers trained on GloVe embedding.

4.4. State-of-the-Art Comparison

Finally, we conducted a state-of-the-art comparison between our results and those of prior research works conducted on drug SAs. As discussed in the Related Works Section, most of the research works utilized DL classifiers to obtain a great accuracy score. Additionally, their studies lacked significant classification scores using ML classifiers, which

are computationally fast compared to DL models. In this study, we outperformed the results of previous research works [14,18,19] on multiclass drug SAs conducted on a drugs.com dataset. Table 4 indicates that with the use of ML algorithms, we obtained a better accuracy score than prior works on DL algorithms.

Table 4. State-of-the-art comparison.

Reference	Dataset	Approach	Number of Classes	Accuracy
Our Approach	**Drugs.com**	**ML**	**3**	**96.60%**
		DL		**97.4%**
[14]	Drugs.com	ML	3	93.80%
[18]	Drugs.com	ML	3	93.85%
[19]	Drugs.com	DL	3	90.40%

4.5. Web Application Development

We also built a web-application-based automatic drug review categorization tool using the highest performing ML model. The Flask framework was utilized to create the application due to its scalability, lightweight features, and availability of Python libraries for the task of conducting the SA. Figure 14 shows the results page of the web application, where users could provide reviews of a drug and the count-vectorized RF model predicted the text as a neutral class.
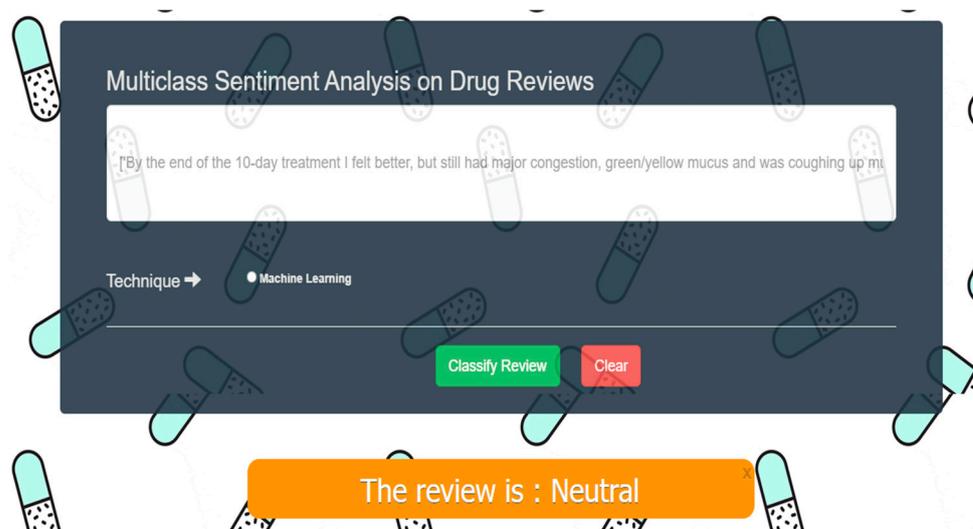

Figure 14. Example of drug review sentiment analyzer web application.

5. Discussion

This paper presented a comparative performance analysis of ML/DL algorithms trained on different feature extraction and word-embedding techniques to classify drug sentiments into three classes: positive, neutral, and negative. Since the dataset used in this study had a lower number of neutral reviews, the ML models encountered difficulties to predict the classes accurately. On the other hand, the DL algorithms trained on GloVe predicted the neutral class more accurately, exceeding the accuracy scores of the ML models. Though we obtained better results than previous research works, the scores could still be improved by increasing the neutral reviews in the dataset.

Previously, much research was conducted on drug SAs using both ML and DL algorithms, where DL classifiers resulted in significant accuracy scores, but ML models suffered from poor accuracy scores. Therefore, in this article, we first focused on exceeding the classification score of previous research works on drug SAs using only ML algorithms com-

bined with several NLP techniques. Out of the two feature extractors, the RF model trained on a CV obtained the highest accuracy and F1 score of 96.65% and 96.42%, respectively. Our RF model outperformed the highest performing DL algorithms of previous research works on drug SAs [18]. Text classification presented challenges that are well suited for RF classifiers due to the large dimensionality and noise of the data. Since the dataset we used in this experiment was imbalanced, a greater classification score could be achieved with RF than with other methods thanks to the majority vote on the predictions determined by all of the decision trees in the forest. It is also worth mentioning that all the classifiers trained with the CV performed equally well and outperformed the results of prior research works. The utilization of several preprocessing techniques cleaned the reviews and reduced the dimensionality of the data. As a result, with the reduced dimension of data, all the ML algorithms reached a significant accuracy score. In addition, throughout the research, we found out that models trained with the CV could obtain greater accuracy than TF–IDF-vectorized models for the task of conducting a multiclass sentiment classification of drug reviews. The more straightforward and honest manner in which the CV represented the words of the evaluations was the primary contributor to its superior performance. While the CV mostly indicated the total number of words that were included in the reviews, TF–IDF primarily represented the importance of the words that were present in the reviews. The TF–IDF encoding performed worse than the CV encoding because it concealed the complete contextual etymology of the text inside the review.

Both the Bi-LSTM and LSTM models trained on GloVe embedding outperformed the performance of the count-vectorized RF, with an accuracy and F1 score of over 97%. The Bi-LSTM model could process inputs both forward and backward in time due to the Bi-LSTM processing chain, replicating the LSTM processing chain. By including a second hidden layer, Bi-LSTM improved the unidirectional LSTM by enabling hidden-to-hidden interconnections to transmit in the opposing temporal sequence. As a result, the model could use data from both the present and the future. For sentiment classification issues, it is beneficial for a model to be aware of both past and future contexts. The method enables Bi-LSTM to take the future context into account. Additionally, without keeping duplicate context information, its layer learns bidirectional long-term dependency between time steps in time series or sequence data. When we wanted the network to learn from the full time series at each time step, while simultaneously having access to contextual data, these dependencies were essential. As a result, it proved to have excellent performance in our study. However, compared to other classifiers, the Bi-LSTM model had the drawback of requiring more training data and time.

6. Conclusions

It is of high significance to examine the feelings of reviews via the use of AI technology in our day and age, characterized by the fast growth of the Internet technology and social networks. When we go out to accomplish anything, be it go shopping, carry out an online purchase, or go to a restaurant, we first look at the reviews to ensure that we are making the best choice. Reviews are becoming an increasingly important part of our everyday lives. The utilization of drug reviews can shed light on the knowledge of users' preferences and drug experiences, which can be exploited to assist healthcare experts with decision making and promote health. SA on online drug reviews is subject to various limitations and challenges, including the presence of subjective and conflicting opinions, the complexity of human language, and the potential for bias in the training data and models. ML/DL algorithms have the ability to provide a fast and straight-forward way of understanding and evaluating online drug reviews. Our study provided a comparative performance analysis of ML/DL algorithms, trained on different feature extraction techniques, seeking to determine the underlying attitude conveyed within online drug reviews with improved accuracy scores from prior studies.

Previous research studies on drug SAs using ML techniques suffered from poor accuracy scores. The accurate determination of sentiments expressed in drug reviews

is crucial for improving patient care, drug development, drug discovery processes, and monitoring public perceptions. As a result, our research aimed to address this issue by studying SA on drug reviews using various ML classifiers, such as RF, SVM, LR, PAG, and SGD, as well as DL classifiers, such as LSTM, Bi-LSTM, GRU, and Bi-GRU. Firstly, we trained the ML models using two feature extraction techniques, namely, TF–IDF and CV, to determine which algorithms could provide better classification scores. Secondly, we utilized two distinct word-embedding techniques, word2sequence and GloVe embedding, to train the DL classifiers in order to further improve the results and overcome the limitations of the ML models. We evaluated the performance of the algorithms using traditional metrics, such as accuracy, precision, recall, and F1 score, and also conducted an error analysis using a confusion matrix and ROC–AUC curve. Our results demonstrated that, among the ML algorithms, RF trained on the CV achieved the highest accuracy score of 96%, surpassing the results of previous research studies. Additionally, models trained with the CV outperformed the results of the TF–IDF-vectorized model due to their straightforward approach in representing the words of the reviews. For the DL classifiers, the Bi-LSTM model trained on GloVe outperformed the ML classifiers with an accuracy and F1 score of 97.40% and 97.42%, respectively. Bi-LSTM performed better due to its ability to extract relevant information from lengthy reviews more effectively by dealing with forward–backward dependencies from feature sequences and resolving gradient disappearance and long-term dependence. Furthermore, we developed a web application based on the count-vectorized RF model that could automatically categorize drug reviews into three classes.

Our study not only enhanced the classification scores of previous research works, but also introduced a web application capable of identifying drug safety concerns and supporting healthcare professionals in determining informed treatment decisions, thus, offering the potential to improve healthcare. By incorporating scalable and efficient algorithms and language representation models such as BERT, GPT, ULMFiT, and Transformer-XL in our future works, along with variational autoencoders and adversarial networks as part of our semisupervised analysis strategies, we plan to explore other approaches that can reduce the dependence on annotated corpora. Furthermore, we aim to investigate the use of semantic features in addition to contextual ones to increase the level of fine-tuning in our strategies. However, interdisciplinary collaboration among computer scientists, healthcare professionals, and regulators is essential to address these research gaps.

References

1. Atkinson, R.K.; Sabo, K.; Conley, Q. The Participatory Web. In *Handbook of Technology in Psychology, Psychiatry and Neurology: Theory, Research, and Practice*; Nova Science Publishers: Hauppauge, NY, USA, 2012; pp. 91–120. [CrossRef]
2. Raynor, D.; Blenkinsopp, A.; Knapp, P.; Grime, J.; Nicolson, D.; Pollock, K.; Dorer, G.; Gilbody, S.; Dickinson, D.; Maule, A.; et al. A systematic review of quantitative and qualitative research on the role and effectiveness of written information available to patients about individual medicines. *Health Technol. Assess.* **2007**, *11*, 1–160. [CrossRef] [PubMed]
3. Mickan, S.; Tilson, J.K.; Atherton, H.; Roberts, N.W.; Heneghan, C. Evidence of effectiveness of health care professionals using handheld computers: A scoping review of systematic reviews. *J. Med. Internet Res.* **2013**, *15*, e212. [CrossRef] [PubMed]
4. Lee Ventola, C. Social Media and Health Care Professionals: Benefits, Risks, and Best Practices. *Pharm. Ther.* **2014**, *39*, 491. Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103576/ (accessed on 18 February 2023).

5. 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 Distance Education and Learning, Lyon, France, 23–26 April 2018; pp. 121–125. [CrossRef]

6. Chen, T.; Samaranayake, P.; Cen, X.; Qi, M.; Lan, Y.C. The Impact of Online Reviews on Consumers' Purchasing Decisions: Evidence from an Eye-Tracking Study. *Front. Psychol.* **2022**, *13*, 2723. [CrossRef]

7. Wankhade, M.; Rao, A.C.S.; Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. *Artif. Intell. Rev.* **2022**, *55*, 5731–5780. [CrossRef]

8. Jiménez-Zafra, S.M.; Martin-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.* **2019**, *93*, 50–57. [CrossRef] [PubMed]

9. Ordenes, F.V.; Theodoulidis, B.; Burton, J.; Gruber, T.; Zaki, M. Analyzing Customer Experience Feedback Using Text Mining: A Linguistics-Based Approach. *J. Serv. Res.* **2014**, *17*, 278–295. [CrossRef]

10. He, W.; Wu, H.; Yan, G.; Akula, V.; Shen, J. A novel social media competitive analytics framework with sentiment benchmarks. *Inf. Manag.* **2015**, *52*, 801–812. [CrossRef]

11. Haque, R.; Islam, N.; Tasneem, M.; Das, A.K. Multi-class sentiment classification on Bengali social media comments using machine learning. *Int. J. Cogn. Comput. Eng.* **2023**, *4*, 21–35. [CrossRef]

12. Haque, R.; Islam, N.; Islam, M.; Ahsan, M. A Comparative Analysis on Suicidal Ideation Detection Using NLP, Machine, and Deep Learning. *Technologies* **2022**, *10*, 57. [CrossRef]

13. Zapf, D. Emotion work and psychological well-being: A review of the literature and some conceptual considerations. *Hum. Resour. Manag. Rev.* **2002**, *12*, 237–268. [CrossRef]

14. Garg, S. Drug Recommendation System Based on Sentiment Analysis of Drug Reviews Using Machine Learning. In Proceedings of the 11th International Conference on Cloud Computing, Data Science and Engineering, Noida, India, 28–29 January 2021; pp. 175–181. [CrossRef]

15. Uddin, M.N.; Bin Hafiz, F.; Hossain, S.; Islam, S.M.M. Drug Sentiment Analysis using Machine Learning Classifiers. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 92–100. [CrossRef]

16. Chen, T.; Su, P.; Shang, C.; Hill, R.; Zhang, H.; Shen, Q. Sentiment Classification of Drug Reviews Using Fuzzy-rough Feature Selection. In Proceedings of the IEEE International Conference on Fuzzy Systems, New Orleans, LA, USA, 23–26 June 2019. [CrossRef]

17. Vijayaraghavan, S.; Basu, D. Sentiment Analysis in Drug Reviews using Supervised Machine Learning Algorithms. 2020. Available online: https://arxiv.org/abs/2003.11643v1 (accessed on 4 June 2022).

18. Colón-Ruiz, C.; Segura-Bedmar, I. Comparing deep learning architectures for sentiment analysis on drug reviews. *J. Biomed. Inform.* **2020**, *110*, 103539. [CrossRef]

19. Beam, A.L.; Kohane, I.S. Big Data and Machine Learning in Health Care. *JAMA* **2018**, *319*, 1317–1318. [CrossRef]

20. Taherdoost, H.; Madanchian, M. Artificial Intelligence and Sentiment Analysis: A Review in Competitive Research. *Computers* **2023**, *12*, 37. [CrossRef]

21. Na, J.-C.; Kyaing, W.Y.M. Sentiment Analysis of User-Generated Content on Drug Review Websites. *J. Inf. Sci. Theory Pract.* **2015**, *3*, 6–23. [CrossRef]

22. Korkontzelos, I.; Nikfarjam, A.; Shardlow, M.; Sarker, A.; Ananiadou, S.; Gonzalez, G.H. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. *J. Biomed. Inform.* **2016**, *62*, 148–158. [CrossRef]

23. Chang, C.; Masterson, M. Using Word Order in Political Text Classification with Long Short-term Memory Models. *Politi. Anal.* **2020**, *28*, 395–411. [CrossRef]

24. Patel, A.; Meehan, K. Fake News Detection on Reddit Utilising CountVectorizer and Term Frequency-Inverse Document Frequency with Logistic Regression, MultinominalNB and Support Vector Machine. In Proceedings of the 2021 32nd Irish Signals and Systems Conference, ISSC 2021, Athlone, Ireland, 10–11 June 2021. [CrossRef]

25. Saputri, Y.R.; Februariyanti, H. Sentiment analysis on shopee e-commerce using the naïve bayes classifier algorithm. *J. Mantik* **2022**, *6*, 1349–1357. Available online: https://iocscience.org/ejournal/index.php/mantik/article/view/2397/2012 (accessed on 1 March 2023).

26. Singh, K.N.; Devi, S.D.; Devi, H.M.; Mahanta, A.K. A novel approach for dimension reduction using word embedding: An enhanced text classification approach. *Int. J. Inf. Manag. Data Insights* **2022**, *2*, 100061. [CrossRef]

27. Shi, Y.; Yang, Y.; Liu, Y. Word Embedding Representation with Synthetic Position and Context Information for Relation Extraction. In Proceedings of the 9th IEEE International Conference on Big Knowledge, ICBK 2018, Singapore, 17–18 November 2018; pp. 106–112. [CrossRef]

28. Mansour, Y.; Schain, M. Learning with Maximum-Entropy Distributions. *Mach. Learn.* **2001**, *45*, 123–145. [CrossRef]

29. Islam, Z.; Liu, J.; Li, J.; Liu, L.; Kang, W. A semantics Aware Random Forest for Text Classification. In Proceedings of the International Conference on Information and Knowledge Management, Beijing, China, 3–7 November 2019; pp. 1061–1070. [CrossRef]

30. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22. Available online: https://cogns.northwestern.edu/cbmg/LiawAndWiener2002.pdf (accessed on 5 June 2022).

31. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

32. Shi, D.; Yang, X. Support Vector Machines for Land Cover Mapping from Remote Sensor Imagery. In *Monitoring and Modeling of Global Changes: A Geomatics Perspective*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 265–279. [CrossRef]

33. Sharma, U.; Saran, S.; Patil, S.M. Fake News Detection using Machine Learning Algorithms. *Int. J. Eng. Res. Technol.* **2021**, *9*, 509–518. Available online: https://iopscience.iop.org/article/10.1088/1757-899X/1099/1/012040/pdf (accessed on 1 March 2023).

34. Antoun, W.; Baly, F.; Hajj, H. AraBERT: Transformer-Based Model for Arabic Language Understanding. 2020. Available online: http://arxiv.org/abs/2003.00104 (accessed on 5 June 2022).

35. Huang, J. Detecting Fake News with Machine Learning. *J. Phys. Conf. Ser.* **2020**, *1693*, 012158. [CrossRef]

36. El Bilali, A.; Taleb, A.; Nafii, A.; Alabjah, B.; Mazigh, N. Prediction of sodium adsorption ratio and chloride concentration in a coastal aquifer under seawater intrusion using machine learning models. *Environ. Technol. Innov.* **2021**, *23*, 101641. [CrossRef]

37. Wang, X.; Jing, L.; Lyu, Y.; Guo, M.; Wang, J.; Liu, H.; Yu, J.; Zeng, T. Deep Generative Mixture Model for Robust Imbalance Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 2897–2912. [CrossRef]

38. Zhang, J.; Zeng, Y.; Starly, B. Recurrent neural networks with long term temporal dependencies in machine tool wear diagnosis and prognosis. *SN Appl. Sci.* **2021**, *3*, 442. [CrossRef]

39. Zulqarnain, M.; Ghazali, R.; Hassim, Y.M.M.; Rehan, M. Text classification based on gated recurrent unit combines with support vector machine. *Int. J. Electr. Comput. Eng.* **2020**, *10*, 3734–3742. [CrossRef]

40. Yang, M.; Moon, J.; Yang, S.; Oh, H.; Lee, S.; Kim, Y.; Jeong, J. Design and Implementation of an Explainable Bidirectional LSTM Model Based on Transition System Approach for Cooperative AI-Workers. *Appl. Sci.* **2022**, *12*, 6390. [CrossRef]