International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 / https://doi.org/10.14445/22315381/IJETT-V70I7P208

Volume 70 Issue 7, 68-78, July 2022

Original Article

A New Classifier Model on Drug Reviews Dataset by VADER Sentiment Analyzer to Analyze Reviews of the Dataset are Real or Fake based on Machine Learning

Manish Suyal[1], Parul Goyal[2]

[1,2]*Department of CA & IT, SGRR University, Uttarakhand, India.*

[1]suyal.manish923@gmail.com

Received: 25 May 2022 Revised: 25 June 2022 Accepted: 30 June 2022 Published: 18 July 2022

Abstract *- Machine There was a time when the customer needed direct advertising and word of mouth to choose the right product. Nowadays, the internet makes the same work very easily accessible to many people who want to know what others think of an item before buying it. Apart from this, knowing the real approach of the business towards its product can greatly benefit the business. These days people can express their feelings in many ways, such as Twitter, Facebook or Instagram, blog posts, and reviews websites. People can freely express their views about any product and service by coming on all these platforms. Therefore, a scholar can use sentiment analysis in health-related facilities. The scholar will develop such a classifier model keeping the VADER Sentiment Analyzer of sentiment analysis in mind. People's opinion is very important, and based on people's opinion, business is done nowadays, and people are also being helped by their opinions. Many people express their opinions on online platforms like Facebook and Twitter. Nowadays, people's opinions are needed in every field because business is done. The paper can help people in any field, whether it is the field of business or medicine, or the field of science. The research scholar can apply sentiment analysis to extract important information from them in a hidden form on these opinions. This important information can be very useful for any field such as medicine, business, and other fields. So the research scholar will develop the proposed drugs reviews recommended system model based on the VADER Sentiment Analyzer of sentiment analysis that will analyze the reviews given about the drugs and will tell whether the given reviews are genuine or fake and on the basis, a patient will be recommended drugs through the proposed model.*

Keywords *– Artificial Intelligence, Machine learning, Sentiment Analysis, VADER Sentiment Analyzer, Opinion Mining, Confusion Matrix, Polarity Scores, Supervised Learning, Unsupervised Learning.*

1. Introduction

In today's environment, sentiment analysis is most commonly used on customer data. The important information can be retrieved from a large amount of data in a very short time. It helps a lot for companies who always want their brand to be viewed from a positive perspective or to have better responses and reactions in the market. In today's era, companies want to know the customer's feedback to increase the sale of their product reviewed online by the customers. The customers express their feedback or reviews about those products online based on hidden sentiments or meaningful aspects. Companies decide whether to launch their products further in the future or reduce their sales based on the customer's opinions. Nowadays, applications built from the concept of sentiment analysis are monitoring and analyzing the sentiments of people's posts on social media. The result is announced before the elections based on sentiments extracted from these tweets.

The scholar can predict in advance which party will win the election in the future based on the feelings hidden in the people's opinions before the election is held. When sentiment analysis is applied to a text, then the process of extracting sentiment from that time is called polarity detection. The polarity detection process of sentiment analysis identifies the hidden sentiment in the text and classifies its text as positive, negative, and neutral. In today's time, many social media posts often express very honest opinions about the company's products and services. With sentiment analysis, the scholar can extract personal emotions from the honest opinions people expressed on social media platforms in a very short time. For the growth of any business, the customer's opinion about the product of that company increase on this matter. The sentiment analysis uses a combination of distinct statistics, machine learning, and natural language processing to identify the subjective information from the text files, such as feelings, decisions, and opinions, assessments about a particular product or brand. The sentiment analysis is used in many approaches,

like reviewing products and services. Still, sentiment analysis is being used a lot in health care too. Nowadays, most of the health care information is available online. If a patient has to decide about their medical problem, then the experience of other patients is very useful for them. All these decisions can be learned only from the experience of any other patient. The information greatly helps the hospitals know the patients' interests and problems and solve them. The objective of the paper is to develop a proposed drugs reviews recommended system model based on the VADER Sentiment Analyzer of sentiment analysis that will analyze the reviews given about the drugs and will tell whether the given reviews are genuine or fake and based on a patient will be recommended drugs through the proposed model.

This model helps those patients who live in remote places where hospitals and medical facilities are not available. The proposed model directs needy patients to the right drug to treat any disease, where no doctor and medical facility is available offline. In addition, the scholar can find the performance of the proposed model by the confusion matrix in the form of accuracy, precision, recall, and f1-score values. With this, the scholar will also compare the proposed classifier model's performance with the previous classifier model used in previous years.

- The classification was one of the hallmarks of all the machine learning models scholars used over the years. This scholar used to see that when the scholar had data in the form of supervised learning, there was a value in the form of dependent and independent variables. The scholar used the machine learning model and looked at the label data by classifying it.
- But as scholars move towards sentiment analysis, the concept comes in the form that scholars will have the same data, but the data will not be labeled. This scholar has to use the pre-trained library already in python language and by which the scholar will divide the existing data into three parts. After this extracting its compound score and presenting its output in binary form.
- The scholar can understand the in such a way that when scholar can try to understand a sentence, there are three types based on sentiment analysis. It can be positive or negative, and it can be a neutral sentence.
- These terms (Positive, Negative, and Neutral) have a separate probabilistic value based on which the scholar will keep a compound value. On the compound, the score scholar will put a threshold value.
- Based on which this comment and sentiment can be positive or negative and can be either neutral.
- This scholar will use the Natural Language Toolkit (NLTK'S), inside which a module is present by the name of the VADER module.
- The VADER module can show the raw positive, negative, and neutral data.
- VADER is a module of natural language processing (NLTK'S), with the help of which it counts the sentiment score. The scholar has already expressed the calculation of sentiment scores in three forms positive, negative or neutral.
- VADER stands for Valence Aware Dictionary for Sentiment Reasoning. VADER is an NLTK's (Natural Language Processing Tool Kit) that provides Sentiment scores based on the words.
- VADER (Valence Aware Dictionary and Sentiment Reasoner) is a lexicon that scholars can consider as a module of rule-based sentiment analysis.
- VADER module makes great use of the sentiment lexicon combination with another large list of linguistic features such as words on which words are usually labeled as positive, negative, and neutral according to their semantic orientation.
- In this way, the scholar can understand that VADER is a type of sentiment analysis module that relies on a dictionary of sentiment-related words. In this module, each word in the vocabulary is assessed as having a positive, negative, or neutral context. An example can be seen in the VADER module, where more positive words have a higher positive evaluation and more adverse words have lower negative grades.

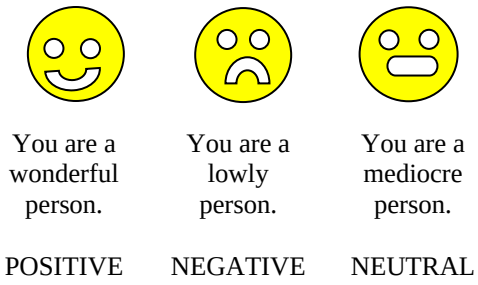


Fig. 1 Sentiment Analysis

2. Related Work

The research scholar has studied the review papers of the last 10 years to understand the work done in the previous year related to the unsupervised machine learning algorithm.

Sahayak V, Shete V, [1] have noticed the paper nowadays; the social networking market has increased significantly. Many people want to know the opinion of people on social networking sites to buy their product. In this paper, the authors discuss a paradigm for extracting sentiment from micro-blogging services like Twitter, where users post their opinions on everything. In this paper, the authors have analyzed the Twitter dataset from a data mining approach using machine learning and sentiment analysis algorithms. The paper states that the proposed sentiment analysis on the Twitter dataset is divided into two important parts, the first one is data extraction, and the second one is preprocessing applied to the extracted dataset and classifying it.

In the paper, Andrea and Ferri [2] convey that working on sentiment analysis is a complex process. Sentiment analysis is increasingly used in machine learning. The different 4 steps to do sentiment analysis are as follows. These steps are collecting the data, preparing the text, detecting the sentiment, and classifying the sentiment.

The information described in the paper is based on the SentiWordNet approach or algorithm, a lexical resource for opinion mining. In the SentiWordNet algorithm, triple polarity is assigned for each synset of the wordnet. That is, positive, negative, and neutrality is assigned to it. The sum of all these scores is always one. For example, {0, 1, 0} are positive, negative, and negative. The SentiWordNet technique is created automatically through linguistic and statistical classifiers. It is very easily implemented in various opinion-related tasks [3].

In this paper, the author explains the sentiment analysis methodology and defines how many steps the work of sentiment analysis is completed. In the first step, do the preprocessing step, in which data is first cleaned to reduce. Then the second stage is the feature extraction, in which a token is assigned to the keyword, and the token is now put under analysis. The third step is the classification in which these search keywords are placed under some category based on machine learning algorithms like K-Nearest Neighbor and Random Forest [4].

Bose and Aithal [5] have analyzed the paper and considered that many people from all over the world set their point of view on important social networking podiums like Twitter these days. In today's environment, from the Twitter dataset, a scholar can get a lot of important information and comments about many products, fashion, etc. accordingly, the scholar can buy that product or that fashion item based on the people's opinion. Using sentiment analysis on the Twitter dataset, scholars can identify people's opinions about the item behind those reviews. In the paper, the sentiment analysis of public opinion about the drugs for the therapy of COVID-19 has been identified. Under that, how many classes of people have benefited and have suffered from COVID-19 disease can be understood. The people have given various reviews about the drugs for the therapy of COVID-19 on the Twitter dataset.

The research scholar tried to identify the sentiment behind it using VADER Sentiment analysis. The Vader sentiment analyzer calculates whether the drug review's polarity is positive, negative, or neutral.

Nowadays, in recent years, business and public sentiments are getting a new look based on the people posting their opinions on social media, and that things are helping businesses a lot. When a scholar can decide on something, the scholar usually wants to know the opinion of others about that thing. Every business and organization wants to know customers' opinions about their products and services. In the article, the scholar has used NLTK, Text blob, and VADER sentiment analysis tool on the movie reviews and classified the movie reviews according to the polarity calculation. The VADER sentiment analyzer finds the polarity calculation in Precision, Recall, F-1 score, and compound value and classifies the reviews as positive, negative and neutral [6].

There are reviews about many things by people on Twitter or other social networking podium websites. To extract the sentiment from it, the scholar can remove those stop words; such words do not play any significant role in sentiment analysis. These words combine prepositions, names, bases, verbs, numbers, etc. [7].

The study [8] describes, In the machine learning field, Naïve Bayes is a potential classifier with a simultaneous assumption which plays an important role in classifying the higher performing classes. The classification of a review can be done easily with the Naïve Bayes machine learning algorithm based on sentiment analysis. Still, Naïve Bayes is a very simple classifier and does not give good results compared to other classifiers such as SVM (Support Vector Machine), DT (Decision Tree), etc.

$$P(X|Y_i) = \pi_{i=1}^{m} P(X|Y_i) \qquad (1)$$

X is defined as feature vector X= $\{x_1, x_2 \ldots x_m\}$ and Y_i is displayed at the class level.

Ahmad, and Aftab [9], explain in the paper that different machine learning tools and techniques have been worked in sentiment analysis and classification. The different famous algorithms like Maximum Ent, Random forest, Naïve Bayes, and Support Vector Machine (SVM) have been discussed in the paper through the sentiment classification can be done easily.

Mahesh [10] analyzed that machine learning algorithms are used extensively for scientific study and statistical models. A computer is explained as an example, then the computer analyzes these examples and automatically proceeds. The most advantage of the machine learning algorithm is that algorithm learns what to do with the data at once, and the model works automatically.

Prakash and Imambi [11] describe that the paper's main objective is to comprehensively review the prediction and evaluation of the COVID-19 datasets based on the opinions given by the people through online mode. The coronavirus belongs to the genus and is a virus that has no vaccine. The coronavirus has wreaked havoc on human life and financial and economic systems in every country worldwide. COVID-19 will stop everything in society. The machine learning

classification algorithm has been applied to the COVID-19 dataset, and scholar is trying to understand which age group is most affected by COVID-19.

Supervised learning is a machine learning algorithm taught to the system from examples. Based on these examples, a computer or model automatically performs further actions and predicts other future examples. This paper compares various supervised learning algorithms and determines the best classification technique for the model. The seven machine learning algorithms were considered like Decision Table (DT), Random Forest (RF), Naive Bayes (NB), Support Vector Machine (SVM), Neural Network, JRip, and Decision Tree(J48) [12].

In today's time, to find the hidden irregularities and regularities in social data, a scholar is using machine learning algorithms to a great extent because, in this way, the scholar can predict anything in the future. This paper attempts to understand the concept and development of some well-known machine learning algorithms. It compares the three algorithms, Decision Tree (DT), Support Vector Machine (SVM), and Bayesian, based on some basic assumptions. The diabetes datasets have been used and compared the performance of each algorithm in terms of training time, prediction time, and prediction accuracy [13].

Navin and Pankja [14] describe nowadays, text mining plays an important role in classifying the text of digital documents. Classifying text in a document requires a lot of techniques such as text mining, natural language processing, and machine learning algorithms. In this paper, a scholar has analyzed the different classification algorithms like K-Nearest Neighbor, Naïve Bayes, and Support Vector Machine. The scholar uses the precision, recall, and f-measure formula, which the scholar gets from the confusion matrix, to calculate the precision, recall, and F-1 score of the model trained with these machine learning algorithms.

In healthcare industries, huge amounts of data are collected but not mined properly, and it doesn't get put to good use. The search for these hidden patterns and relationships often goes unappreciated due to not getting to know much information. The research focuses on developing a prediction model to extract some such information for diagnosing a disease like diabetes, hepatitis and heart diseases, etc. so that a disease can be easily dealt with in the coming times. To find the hidden patterns, scholars need machine learning classification algorithms like Support Vector Machine and Decision Tree [15].

3. Sentiment Analysis VADER Methodology Using NLTK (Machine Learning Approach)

The scholar has defined the following methodology steps to develop the proposed model. The VADER sentiment analyzer and several libraries of python language, including nltk, pandas, numpy, and sklearn, have been used to develop the proposed model.

Step-1: The scholar has a drug review dataset here with people's reviews, and next to it is a label containing the result of what people think about those reviews. That is, the positive and negative sentences in this label are written. The scholar has to check with the implementation model that the reviews given, and the labels given as positive and negative are genuine reviews or written like that.

Step-2: After this scholar will check the dataset's size and will try to know the total review of the dataset. The scholar will also try to find out how many total positive reviews are in the dataset and what negative reviews are there.

Step-3: To remove the null values and blank space from the dataset, the scholar will follow the preprocessing methodology to clean the data. By doing this, the dataset's data is made in the form of a well-organized manner.

Step-4: The scholar checks the polarity scores of the dataset reviews with the help of the VADER sentiment analyzer. First of all, scholars analyze the first index's review and try to calculate its polarity score. The VADER sentiment analyzer module is imported from python's NLTK library. VADER stands for Valence Aware Dictionary for Sentiment reasoning. The VADER is an NLTK's (Natural Language Processing Tool Kit) that provides Sentiment scores based on the words. VADER (Valence Aware Dictionary and Sentiment Reasoner) is a lexicon that scholars can consider as a module of rule-based sentiment analysis. VADER module makes great use of the sentiment lexicon combination with another large list of linguistic features such as words on which words are usually labeled as positive, negative, and neutral according to their semantic orientation.

Step-5: The scholar gets the polarity scores of the first review in the form of negative, positive, and neutral through the VADER sentiment analyzer. The scholar also gets the compound value of the first index review in the dataset.

Step-6: The way scholar found the polarity score of the first review by the VADER sentiment analyzer in step 4. In the same way, scholars will apply the VADER sentiment analyzer to the whole data of the dataset and extract the polarity scores in the form of positive, negative, neutral, and compound values.

Step-7: Extract the compound value from the polarity scores because the compound value gives you your main result.

Step-8: The compound value of the review is in numerical form. The scholar will convert the compound value into positive and negative. Suppose the compound value exceeds

zero (compound value>0). In that case, the scholar will consider it positive, and if the compound value is less than 0 (compound value<0), then the scholar will consider it a negative means. If the compound value is positive, the review is positive, and if the compound value is negative, then the review is positive.

Step-9: The scholar compares the dataset label value based on the supervised learning technique with the compound value. When scholars compare the label next to the review entered in the dataset with the compound value analyzed by the VADER sentiment analyzer, this shows how many reviews in the dataset are genuine and fake. It is not that any person has written about the reviews like this.

Step-10: Now, the scholar will calculate the accuracy of the proposed model with the help of the sklearn library in python. To get the accuracy of the proposed model, the scholar has to divide the total correctly classified value by the total number of samples.

Accuracy of any model = (True Positive + True Negative)/(True Positive + False Negative + True Negative + False Positive)

accuracy_score(df[‘label’], df[‘comp_score’]) (2)

In python, the scholar can derive the accuracy of any model from a sklearn library using the accuracy_score() function. After analyzing the review and its compound score field, this scholar has given the actual label field and passed it to the accuracy_scores () function, which calculates the model's accuracy. From the accuracy, scholars know how many percent correct prediction models have been made and how many wrong predictions have been made. It is shown in percentage.

Step-11: Now scholar will generate the classification report of the model using the classification_report() function. In this also scholar has to take the label and comp_score.

classification_report (df [‘label’], df [‘comp_score’])) (3)

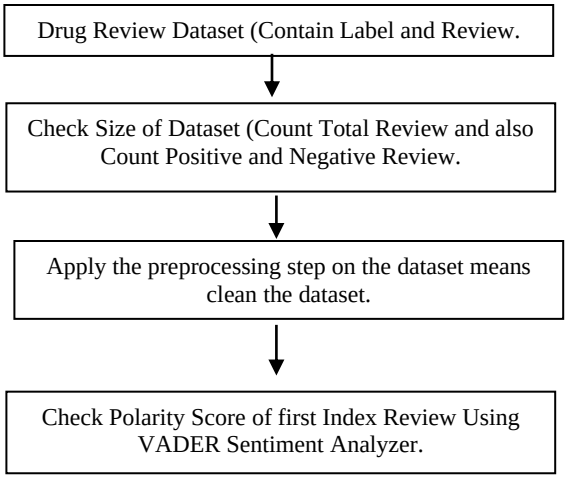
The scholar gets to know the value of precision, recall, and f-measure for the negative and positive reviews in the dataset through the classification report, and the scholar can calculate the average value of the proposed model in terms of precision, recall and f1-score and together scholar get to know how many positive reviews and negative reviews scholar has.

Step-12: Now, in the last, the scholar will get the confusion matrix of the proposed predictive model. The value of the confusion matrix diagonal tells us how many predictions a scholar has made are correct and how many predictions a scholar has made wrong. To do this scholar has to write the code to generate the confusion matrix in python. In what the scholar has written below, a label and compound score must be given to each review.

confusion_matrix (df [‘label’], df [‘comp_score’])) (4)

Step-13: The scholar compared the proposed model accuracy based on the sentiment analysis with the classifier model used in previous years, such as support vector machine (SVM) and Back Propagation (BP). The scholar expects that, after comparing, the proposed model's performance will be much better than the classifier model used in previous years in terms of accuracy, precision, recall, and f1-score. Because in the models used in previous years, the dataset from which the model was trained, the reviews given by the people in it, many times there was a label next to the reviews by the people which was the target attribute if a person writes positive and negative like this by mistake, then if scholar train the classifier model from this dataset, due to this, the classifier model can sometimes make wrong predictions because it is based on supervised learning, but the proposed model classifies the reviews based on sentiment analysis as polarity scores like positive, negative and neutral scores values [16]. For this reason, whatever reviews people give them, and this investigation is done correctly means whatever review has been given by any person. On the same basis, sentiment analysis will analyze it and determine whether those reviews are positive or negative [17].

The scholar has proposed a model based on sentiment analysis; the compound value from this is what the scholar has given labels in the dataset. Sometimes, people have given it like this and compare it with their labels. When the scholar does this, the scholar can find out from the proposed model how many reviews were genuine and fake. By doing this, any needy patient, if they want to know about a drug or medicine in an emergency, can easily get information about the right drug or medicine in the right way and also can use it.



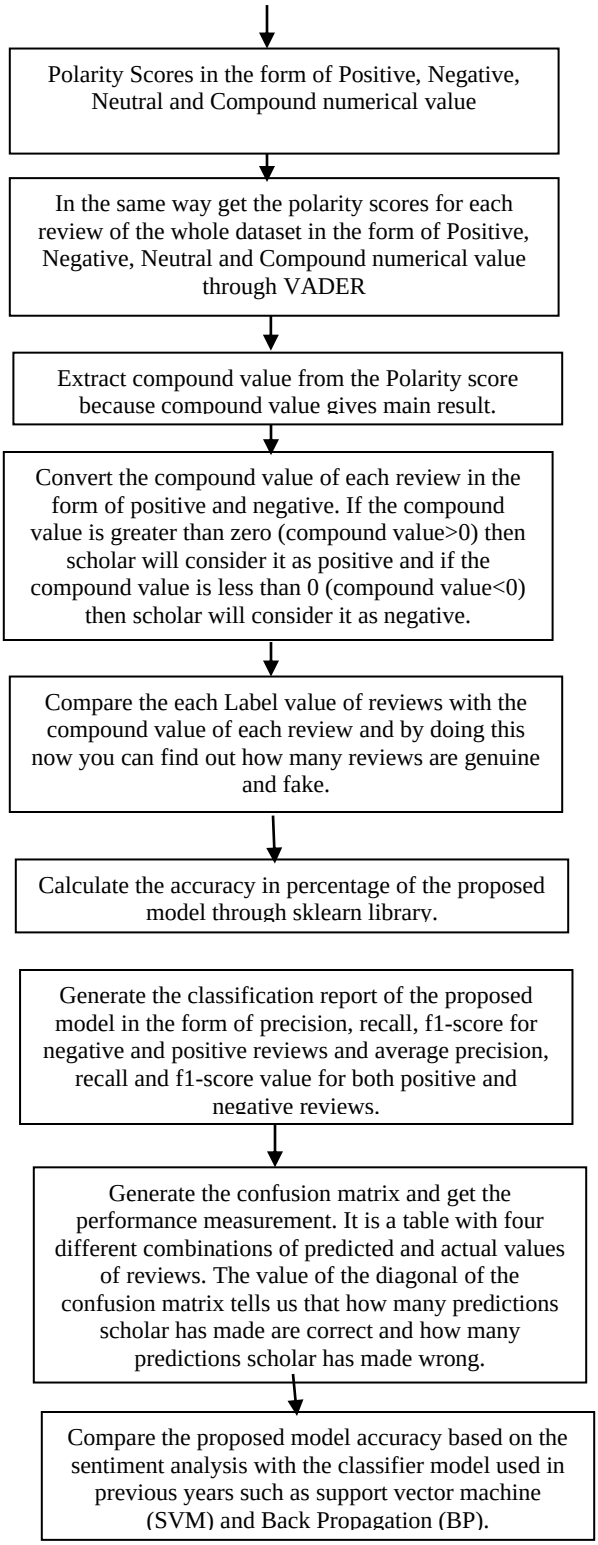


Fig. 2 Problem Design and Methodology

4. Experiment and Evaluation

The scholar can understand the result and evaluation of the proposed Implementation model as follows.

Step-1: The first scholar will do sentiment analysis on a real drug review dataset and use the Bactrim drug review dataset. In this dataset, a scholar has put the review given by the experience patients about the drug. For this, let's extract the dataset. The first scholar will import the entire library.

```
Import nltk
Import numpy as np
Import pandas as pd
```

When the scholar writes this kind of code in python, all the

Libraries will be imported in this way.

Step-2: After that scholar will import the VEDER sentiment analyzer module with the help of NLTK.

```
nltk.download('vader_lexicon')
```

Step-3: In Python, VADER sentiment analysis requires an intensity analyzer function, which scholar will import from

```
nltk.sentiment.vader import SentimentIntensityAnalyzer
sid=SentimentIntensityAnalyzer ()
```

Step-4: Now scholar will take the next level reading the drug review dataset.

```
df=pd.read_CSV ('E:/NLP/TextFiles/Bactrimreviews.tsv, sep='\t')
df.head()
```

Table 1. Drug Review Dataset based on supervised learning

Label	Review
pos	Bactrim drug is very useful for cancer treatment and gives good experience to the patients.
pos	Many people have used Bactrim drug in backward areas, and the people here have found its experience good, and some people have found it normal.
pos	People living in remote areas struggle with drugs when battling cancer, but the Bactrim drug is very useful for cancer. It has brought great relief to the people.
pos	People do not have any side effects from using Bactrim drug, and they get very good relief in treatment like cancer
pos	According to people's opinion for a disease like cancer, Bactrim is very good and useful medicine.

Here, the scholar has a drug reviews dataset, which consists of two things, the review is about the drug and its labels. The scholar will analyze here whether the given reviews are there and whether it is a wrong reviews. These labels should match the analysis. That is, if a positive review has been given here, then it is not that it is a fake review.

Step-5: First, scholars see how big the dataset is and how many reviews and labels are there.
df [‘label’].value_counts()

Table 2. Drug Review Dataset Length in terms of the positive and negative review

neg	5097
pos	4903
name	label, dtype:int64

The table shows 10,000 data points in the dataset, with 5097 negative and 4903 positive reviews.

Step-6: Now scholar will do the preprocessing of the dataset here means the data will be cleaned. If there is a null value, the scholar will remove it, and also, if there is any blank space, then remove it too.
df.dropna(inplace=True)
empty=[]
for i,lb,rv in df.itertuples():
if type(rv)==str;
if rv.isspace():
empty.append(i);

The preprocessing step is complete, and the dataset's data is clean. The blank space and null values have been removed from the dataset, and the whole data has been cleaned.

Step-7: First, the scholar will check the review's polarity above the first index's position. The scholar will check the label of that review.
sid.polarity_scores(df.loc[0][‘review’])
output: {‘neg’:0.088, ‘neu’: 0.669, ‘pos’: 0.243, ‘compound: 0.9454}
If i process df. loc[0][‘review’], the first review will come here.
Output: ‘stunning even for the non-gamer: This soundtrack was beautiful.’

It has a little negative glimpse, but this review is positive overall. The maximum portion inside it is neutral, the positive impact is 0.243, and where the value of the compound is coming from is highly positive. Overall, it is coming out that this is a positive response, and if you look individually at the dataset, this review’s label also looks positive. It means that the review is genuine and not written like this.

Step 8: Now that scholar has one label of data, they can apply it to the whole data in the dataset. This scholar will apply a function to the whole dataset, checking all the reviews in one go. The scholar will collect all the scores in one go, which will apply a lambda function here, which will do all the datasets in one go.
df[‘scores’]=df[‘review’].apply(lambda review:sid.ploarity_scores(review))
df.head()

Table 3. Polarity values of each review in the drug review dataset

Label	Review	Scores
pos	Bactrim drug is very useful for cancer treatment and gives good experience to the patients.	{neg: 0.088, ‘neu’:0.669, ‘pos’:0.243}
pos	Many people have used Bactrim drug in backward areas, and the people here have found its experience good, and some people have found it normal.	{neg: 0.018, ‘neu’:0.837, ‘pos’:0.145}
pos	People living in remote areas struggle with drugs when battling cancer, but the Bactrim drug is very useful for cancer. It has brought great relief to the people.	{neg: 0.04, ‘neu’:0.669, ‘pos’:0.268}
pos	People do not have any side effects from using Bactrim drug, and they get very good relief in treatment like cancer	{neg: 0.09, ‘neu’:0.615, ‘pos’:0.295}
pos	According to people's opinion for a disease like cancer, Bactrim is very good and useful medicine.	{neg: 0.00, ‘neu’:0.746, ‘pos’:0.245}

Step-9: Now all the scores have come out here because if the scholar has the scores, so all the compound value scholar has, that’s here scholar can get it. The scores scholar will extract the compound value, giving you the main result. It will be found from the compound value only.

df[‘compound’]=df[‘scores’].apply(lambda-score_dict: score_dict[‘compound’])

df.head()

Table 4. Polarity values (scores values) with the compound value of each review in the drug review dataset

Label	Review	Scores	Compound
pos	Bactrim drug is very useful for cancer treatment and gives good experience to the patients.	{neg: 0.088, 'neu':0.669, 'pos':0.243}	0.9454
pos	Many people have used Bactrim drug in backward areas, and the people here have found its experience good, and some people have found it normal.	{neg: 0.018, 'neu':0.837, 'pos':0.145}	0.8957
pos	People living in remote areas struggle with drugs when battling cancer, but the Bactrim drug is very useful for cancer. It has brought great relief to the people.	{neg: 0.04, 'neu':0.669, 'pos':0.268}	0.9858
pos	People do not have any side effects from using Bactrim drug, and they get very good relief in treatment like cancer	{neg: 0.09, 'neu':0.615, 'pos':0.295}	0.9814
pos	According to people's opinion for a disease like cancer, Bactrim is a very good and useful medicine.	{neg: 0.00, 'neu':0.746, 'pos':0.245}	0.9781

Step-10: The scholar has got the value of the compound, so now the scholar can work out the final score here. No scholar has to compare. This is how scholars detect whether a review is positive or negative. The scholar knows this by the compound value. It is positive if the compound value is positive or greater than zero. If it comes in the negative, it is called a negative review. The scholar will do the same check here and put the same condition here.

df['comp_score']=df['compound'].apply(lambda C: 'pos' if C>=0 else 'neg')
df.head()

If the scholar checks now, the scholar will also get the comp_score

The value here is the last. Here it is matching so you can find out how many fake reviews or genuine ones.

Table 5. Comparison of the Comp_Value and Label value of the dataset

Label	Review	Compound	Comp_Value
pos	Bactrim drug is very useful for cancer treatment and gives good experience to the patients.	0.9454	pos
pos	Many people have used Bactrim drug in backward areas, and the people here have found its experience good, and some people have found it normal.	0.8957	pos
pos	People living in remote areas struggle with drugs when battling cancer, but the Bactrim drug is very useful for cancer. It has brought great relief to the people.	0.9858	pos
pos	People do not have any side effects from using Bactrim drug, and they get very good relief in treatment like cancer	0.9814	pos
pos	According to people's opinion for a disease like cancer, Bactrim is very good and useful medicine.	0.9781	pos

Step-11: Now scholar will calculate the accuracy here in the last. The scholar will also generate the classification report here to know how many right and wrong predictions are with the help of a confusion matrix. So the first scholar will import the sklearn library.

From sklearn.metrics import accuracy_score_classification matrix.
accuracy_score(df['label'], df['comp_score'])
output: 0.7091 (accuracy_score)

The accuracy that has come here is 0.7091, and according to the NLP model, this accuracy is good.

Step-12: Now scholar will generate the classification report. In this also scholar has to take the label and comp_score.

print(classification_report(df['label'], df['comp_score']))

Table 6. Classification Report (Precision, Recall, and F1-Score)

	Precision	Recall	F1-score	Support
neg	0.86	0.51	0.64	5097
pos	0.64	0.91	0.75	4903
avg/total	0.75	0.71	0.70	10000

Step-13: Now, in the last, the scholar will get the confusion matrix of the proposed predictive model.
print(confusion_matrix(df[‘label’],df[‘comp_score’]))

Table 7. Confusion Matrix

[[2623	2474]
[435	4468]]

The model accuracy was 70%, which means that the 7000 value is the correct prediction. If the scholar adds the confusion matrix's diagonal value, the scholar will get the correct prediction. For the above matrix, 2623+4468 (first diagonal values) =7091 (right prediction), and if you add the value of the above second diagonal, you will get a wrong prediction. For above matrix 435+2474 (second diagonal values) =2909 (wrong prediction).

Scholars can assume that 3000 is the wrong prediction and 7000 is the right prediction. The scholar compared the proposed model based on the sentiment analysis with the classifier model used in previous years, such as support vector machine (SVM) and Back Propagation (BP). After comparing, the scholar found that the performance of the proposed model is much better than the classifier model used in previous years in the form of accuracy, precision, recall, and f1-score. The specialty of the implementation model presented in the paper is that it is based on unsupervised learning. This scholar is trying to understand the review based on sentiment analysis. The scholar can understand this in such a way that when a scholar tries to understand a sentence, there are three types based on sentiment analysis. It can be positive or negative, and it can be a neutral sentence. This scholar will use the Natural Language Toolkit (NLTK’S), inside which a module is present by the name of the VADER module. The VADER module can show the raw positive, negative, and neutral data. VADER is a module of natural language processing (NLTK’S), with the help of which it counts the sentiment score. The scholar has already expressed the calculation of sentiment scores in three forms positive, negative or neutral. The scholar compared the proposed model based on the sentiment analysis with the classifier model used in previous years, such as support vector machine (SVM) and Back Propagation (BP).

Table 8. Comparison of the predictive model (based on VADER Sentiment Analyzer) with other classifier models used in previous years like SVM and BP

	Accuracy	Precision	Recall	F1-Score
Proposed Model (VADER Sentiment Analyzer)	0.70	0.75	0.71	0.79
Previous Model (Support Vector Machine)	0.67	0.71	0.68	0.75
Previous Model (Back Propagation)	0.64	0.69	0.65	0.73

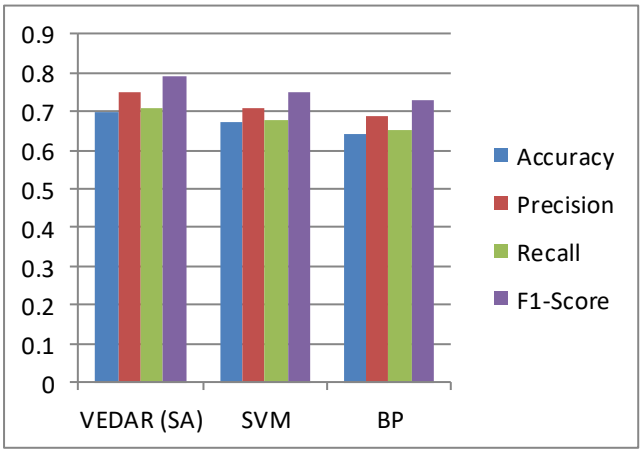


Fig. 2 Comparison Chart of the Proposed Model Performance (VADER Sentiment Analyzer) with the Previous Model (Support Vector Machine and Back Propagation Technique)

5. Conclusion and Suggestions for Future Work

The scholar has presented the implementation model based on sentiment analysis in this paper. Through the proposed model, good medicine will be available in an emergency for many people who live in remote places and where medical and doctor facilities are unavailable. They won’t even have anywhere to go. All these patients can choose their drug by analyzing this proposed model based on the reviews given by the people online on any website or social media platforms like Facebook and Twitter. In addition, scholars find the performance of the proposed model by the confusion matrix in the form of accuracy, precision, recall, and f1-score values. With this, the scholar will also compare the proposed classifier model's performance with the previous classifier model used in

previous years. The scholar found from the implementation of the model that the model based on sentiment analysis is of great importance in properly examining any review. It tells which review is genuine and which is fake. Over the years, many classifier models have garnered online reviews from people. In the model used in the past years, there has been a lot of work on supervised learning, and when the model based on supervised learning is given review data, it requires supervised learning in which you have to give labels next to the review. It often happens when the classifier model is given a label with the review; then, on those labels, people write negative instead of positive in a hurry of their own free will. Due to this, the classifier model gets trained due to some fake reviews. Due to this, the accuracy of the model decreases a lot. The specialty of the implementation model presented in the paper is that it is based on unsupervised learning. This scholar is trying to understand the review based on sentiment analysis. The scholar can understand this in such a way that when a scholar tries to understand a sentence, there are three types based on sentiment analysis. It can be positive or negative, and it can be a neutral sentence.

This scholar will use the Natural Language Toolkit (NLTK'S), inside which a module is present by the name of the VADER module. The VADER module can show the raw positive, negative, and neutral data. VADER is a module of natural language processing (NLTK'S), with the help of which it counts the sentiment score. The scholar has already expressed the calculation of sentiment scores in three forms positive, negative or neutral.

The scholar compared the proposed model based on the sentiment analysis with the classifier model used in previous years, such as support vector machine (SVM) and Back Propagation (BP). After comparing, the scholar found that the performance of the proposed model is much better than the classifier model used in previous years in the form of accuracy, precision, recall, and f1-score. In the future, scholars will work on a much bigger problem inside sentiment analysis. The biggest thing about the reviews is that sarcasm is also added here, which is difficult to recognize. Till now, the library used in python under the natural language processing is also difficult to recognize sarcasm. Sarcasm is such that you use positive words to make negative comments. The meaning is that wording will be positive, but its emotions are negative. For example, consider the sentence: think of yourself as if you have uprooted the mountain.

References

[1] V. Sahayak, V. Shete, "Sentiment Analysis on Twitter Data," *International Journal of Innovative Research in Advanced Engineering (IJRAE)*, vol. 2, no. 1, pp. 178-183, 2015.

[2] A. Dandrea, "Approaches, Tools, and Applications for Sentiment Analysis Implementation," *International Journal of Computer Applications*, vol. 125, no. 3, pp. 0975-8887, 2015.

[3] I. Hemalatha, I. Varma, "Preprocessing the Informal Text for efficient Sentiment Analysis*," International Journal of Emerging Trends and Technology in Computer Science*, vol. 2, no. 1, pp. 58-61, 2012.

[4] P. Baid, N. Chaplot, "Sentiment Analysis of Movie Reviews using Machine Learning Techniques," *International Journal of Computer Applications*, vol. 2, no. 2, pp. 45-49, 2017.

[5] R. Bose, P. Aitha, "Survey of Twitter Viewpoint on Application of Drugs by VADER Sentiment Analysis among Distinct Countries," *International Journal of Management, Technology, and Social Sciences (IJMTS)*, vol. 3, no. 1, pp. 1-18, 2021.

[6] N. Kumaresh, "A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis," *Asian Journal of Computer Science and Technology*, vol. 8, no. 2, pp. 1-6, 2019.

[7] D. Kawade, K. Oza, "Sentiment Analysis: Machine Learning Approach," *International Journal of Engineering and Technology*, vol. 6, no. 2, pp. 2183-2186, 2017.

[8] B. Gupta, F.M. Huber, "Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python," *International Journal of Computer Applications*, vol. 2, no. 6, pp. 29-34, 2017.

[9] M. Ahmed, M. Aftab, "Machine Learning Technique for Sentiment Analysis A Review," *International Journal of Multidisciplinary Science and Engineering*, vol. 8, no. 3, pp. 27-32, 2017.

[10] B. Mahesh, M. Ismail, "Machine Learning Algorithms- A Review," *International Journal of Science and Research*, vol. 9, no. 1, pp. 381-386, 2020.

[11] K. Prakash, S. Imambi, "Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms," *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 5, pp. 2200-2204, 2020.

[12] F.Y Osisanwo, J.E.T. Akinsola, "Supervised Machine Learning Algorithms: Classification and Comparison," *International Journal of Computer Trends and Technology*, vol. 48, no. 3, pp. 128-138, 2017.

[13] K. Dasi, R. Behera, "A Survey on Machine Learning: Concept, Algorithms and Applications," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 5, no. 2, pp. 1301-1309, 2017.

[14] M. Navin, R. Pankaja, "Performance Analysis of Text Classification Algorithms using Confusion Matrix," *International Journal of Engineering and Technical Research*, vol. 6, no. 4, pp. 75-78, 2016.

[15] D. Gillibrand, "The Use of Design Patterns in a Location-Based GPS Application," *International Journal of Computer Science*, vol. 8, no. 3, pp. 1-627, 2011.

[16] M. Suyal, P. Goyal, "An Efficient Classifier Model for Opinion Mining to Analyze Drugs Satisfaction Among Patients," *Communications in Computer and Information Science (CCIS)*, Springer Nature Switzerland AG, vol. 1591, pp. 30-38, 2022. https://doi.org/10.1007/978-3-031-07012-9_3

[17] M. Suyal, P. Goyal, "A Two-Phase Classifier Model for Predicting the Drug Satisfaction of the Patients Based on Their Sentiments," *Communications in Computer and Information Science (CCIS)*, Springer Nature Switzerland AG, vol. 1591, pp. 79–89, 2022. https://doi.org/10.1007/978-3-031-07012-9_7