Intelligent Automation & Soft Computing
DOI: 10.32604/iasc.2022.025861
Article

Tech Science Press

Automatic Annotation Performance of TextBlob and VADER on Covid Vaccination Dataset

Badriya Murdhi Alenzi, Muhammad Badruddin Khan, Mozaherul Hoque Abul Hasanat, Abdul Khader Jilani Saudagar*, Mohammed AlKhathami and Abdullah AlTameem

Information Systems Department, College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, 11432, Saudi Arabia
*Corresponding Author: Abdul Khader Jilani Saudagar. Email: aksaudagar@imamu.edu.sa
Received: 07 December 2021; Accepted: 20 January 2022

Abstract: With the recent boom in the corpus size of sentiment analysis tasks, automatic annotation is poised to be a necessary alternative to manual annotation for generating ground truth dataset labels. This article aims to investigate and validate the performance of two widely used lexicon-based automatic annotation approaches, TextBlob and Valence Aware Dictionary and Sentiment Reasoner (VADER), by comparing them with manual annotation. The dataset of 5402 Arabic tweets was annotated manually, containing 3124 positive tweets, 1463 negative tweets, and 815 neutral tweets. The tweets were translated into English so that TextBlob and VADER could be used for their annotation. TextBlob and VADER automatically classified the tweets to positive, negative, and neutral sentiments and compared them with manual annotation. This study shows that automatic annotation cannot be trusted as the gold standard for annotation. In addition, the study discussed many drawbacks and limitations of automatic annotation using lexicon-based algorithms. The highest level of accuracies of 75% and 70% were achieved by TextBlob and VADER, respectively.

Keywords: Sentiment analysis; lexicon-based approach; VADER; TextBlob; automatic annotation

1 Introduction

Over the past two decades, sentiment analysis or opinion mining has evolved to be a valuable tool in understanding people's emotions with a wide range of usage in various fields such as public health, marketing, sociology, and politics. Creating a ground truth dataset by annotating the data with appropriate sentiment labels indicating positive, negative, and neutral emotion is essential for any sentiment analysis work that uses a supervised learning approach. Traditionally, in supervised sentiment analysis or in general, in any supervised machine learning (ML) approach, dataset annotations are performed by human experts of the respective domain. In sentiment analysis, manual annotations are considered the most accurate reflection of human emotions expressed in any natural language corpus. Hence, manual annotations are the "gold standard" in any sentiment analysis task [1]. The idea that human expert

manual annotations are superior to any other non-human annotation method is essentially an extension of Alan Turing's Turing Test, a fundamental principle in machine learning. The test sets human intelligence as the yardstick to judge the machine's ability to exhibit intelligent behavior. In the context of sentiment analysis, intelligent behavior is the ability to distinguish positive, negative, and neutral emotions in natural language corpora [2].

Undoubtedly, manual annotation of natural language corpora is expensive, time-consuming, and requires domain expertise. These problems are further increased by the sheer increase in data volume and the challenge of real-time analytics demanded by big data applications such as sentiment analysis on Twitter. Some researchers proposed using lexicon-based automatic annotation to tackle these challenges as a replacement for manual annotation [3]. While these approaches may give acceptable solutions in confined experimental settings, it is premature to argue that automatic annotation can outperform manual annotation. VADER [4] and TextBlob [5] are two automatic annotation approaches introduced as alternatives to annotation by human experts. This research presents empirical results showing that while VADER and TextBlob can produce good results, but not to the extent that they can become alternatives to manual annotation.

The effectiveness of the lexicon-based automatic annotation approach employing VADER and TextBlob was tested against manually annotated Arabic tweets about Covid-19 vaccinations in this study. Because TextBlob and VADER are not built for Arabic content, the tweets were manually translated into English using Google Translate and checked for accuracy. Then, for automatic annotations, we used VADER and TextBlob and compared their results to our manual annotations [5].

The main contribution of this research is the creation of a dataset of 5402 tweets about COVID-19 vaccines that were manually annotated as positive, negative, or neutral by native Arabic annotators, followed by the use of manual data annotation to assess the performance and issues with automatic data annotation using VADER and TextBlob.

The paper is organized as follows: Section (2) presents the literature review related to the topic of the presented work. Section (3) shows the methodology used to develop this paper. Section (4) presents the main results of the study. Conclusions and future work are detailed in Section 5.

2 Literature Review

With the rapid evolution and increasingly growing social media on the web, there is a need to analyze public opinions and emotions for better decision-making. Twitter is one of the essential social media platforms. It is a microblog where millions of users succinctly share their ideas and opinions. A significant number of tweets is the input source for big data applications [6] like sentiment analysis.

Big Data is a collection of organized, unstructured, and semi-structured data that is enormous in volume, snowballing, and complicated, making it difficult to process using traditional methods [7]. However, Big Data brings new opportunities to modern society that were impossible with small-scale data. The rapid-growing unstructured posts generated by Twitter users require sentiment analysis tools to discover the sentiments automatically [8]. Sentiment analysis is one of the most critical research areas in natural language processing. It is a field of study that analyzes people's attitudes, opinions, sentiments, and emotions. The growing importance of sentiment analysis is due to the growth of social media like Twitter, blogs, forum discussions, and social networks. Sentiment analysis systems are required in every business and social domain because opinions influence people's behaviors. People's choices and decisions are primarily influenced by how others view the world. Therefore when faced with a decision, they frequently seek out the advice of others [9].

Sentiment analysis requires data annotation representing emotions or sentiments. The annotation process can be manual, or it can be automatic. It can be performed in an automated manner by using two approaches: The lexicon-based (Rule-based) approach and Machine Learning-based approach [10].

2.1 Manual Annotation

Data annotation makes text, audio, or speech understandable through sentiment analysis models. Data annotation is the process of labeling the data available in various formats like text, audio, video, or images to enable machines to quickly and clearly understand and use it. In short, data annotation helps the machines to learn from it to arrive at desired outputs [11]. Manual annotation is the process of assigning labels to blocks of text: whether they are short, long sentences, or paragraphs by a human. Manual annotation is still considered the bottleneck for various Natural Language Processing (NLP) experiments. It is a time-consuming process. It involves various activities, like defining an annotation schema, specifying annotation guidelines, and training experts for the annotation process to build a consensus corpus [12].

2.2 Automatic Annotation

Automatic annotation is the use of sentiment analysis methods, either machine learning approach or lexical-based approach to find patterns in data and discover the emotion of the given information and classify it into one of three classes positive, neutral, and negative

2.2.1 Lexicon-based Approach for Automatic Sentiment Annotation

The text is analyzed without training or applying machine learning models. The result of the analysis is the automatic classification of the text as positive, negative, and neutral. Also, the rule-based approach is known as the lexicon-based approach. Examples of rule-based or lexicon-based approaches are TextBlob and VADER [4].

1) TextBlob

TextBlob is a Python library used for processing textual data. It is an open-source framework with a consistent API for NLP tasks such as noun phrase extraction, part-of-speech tagging, sentiment analysis, classification, and translation. It measures sentiment polarity and subjectivity. Polarity determines if the orientation of the expressed sentiment is positive, negative, or neutral. It is a floating-point number between [−1, 1], +1 indicates extreme positive sentiments, and −1 indicates extreme negative sentiment. Subjectivity determines if the statement relies on beliefs, opinions, assumptions and is influenced by emotions and personal feelings. It is a floating-point number in the range of [0, 1] [5].

2) VADER

It stands for Valence Aware Dictionary and Sentiment Reasoner [4]. It is a model used for text sentiment analysis. It calculates both polarity and intensity of emotion. It is available in the Natural Language Toolkit (NLTK) package. It uses a dictionary to map lexical features to emotion intensities. It gives sentiment scores, which are obtained by summing up the score of each word in the text. The total of positive, negative, and neutral intensities should be 1. The compound score is the metric used to give the overall sentiment; it ranges from −1 to 1. The sentiment is positive if the compound score is greater than or equal to 0.05, neutral if the compound score is between [−0.05, 0.05], and negative if the compound score is less than or equal to −0.05 [4].

2.2.2 Machine Learning Approach for Automatic Sentiment Annotation

It is a supervised learning approach. In this approach, the machine learning classifiers like Support Vector Machine (SVM), Decision Tree (D-Tree), K-Nearest Neighbor (KNN), Naïve Bayes (NB), etc., are used to classify dataset into different sentiments. The approach uses a human-annotated dataset to

build a model that can classify or annotate a large dataset with considerable accuracy. Usually, the accuracy of the machine learning (supervised) approach is higher than the accuracy of the rule-based (unsupervised) approach for sentiment analysis [13].

2.3 Usage of Various Methods for Annotation in Literature

Many studies used lexical-based approaches or machine learning approaches for automatic sentiment analysis. For example, [14] analyzed people's views who interact and share social media like Twitter to present their views regarding COVID-19. The platform for the study experiments was a large-scale dataset COVIDSENTI that consists of 90 000 tweets related to COVID-19 and was collected in the pandemic. The TextBlob tool was used for labeling the sentiment of the tweets into positive, negative, and neutral classes. Then the annotated tweets were used for sentiment classification using different sets of features and classifiers. [15] aimed to improve healthcare services by using lexical-based sentiment analysis approaches to classify patients' opinions. The study found that the accuracy was insufficient to measure the model's performance. Based on precision, recall, and F1-score, the study concluded that VADER lexicon-based approach outperformed the TextBlob model. [16] aimed to build a model reflecting on the sentiment analysis using NLP and different machine learning classifiers like decision trees, random forest, k-nearest neighbors, and Gaussian naïve Bayes. The dataset of the study consisted of 2500 tweets. TextBlob was used to assign labels as positive, negative, and neutral by evaluating the polarity of the dataset. Then the annotated 2500 tweets of the dataset were further divided into testing and training datasets. The study found that the Random Forest Classifier was the most accurate model. [17] presented a hybrid approach that performed analysis and classification of students' feedback before and after the COVID-19 pandemic using ML techniques. Data was collected using a learning management system (LMS), online google forms, and WhatsApp group messages of specific courses. TextBlob and VADER were used to automatically annotate students' feedback on classes before and during the pandemic. Naïve Bayes and Support Vector Machine algorithms were used for classification and comparative analysis. The study achieved an average accuracy of 85.62% by the Support vector machine algorithm.

On the contrary, some studies criticize the performance of automatic annotation for sentiment analysis. For example, [3] provided a detailed comparison of sentiment analysis methods and insisted on the importance of validating automatic text analysis methods using manual annotation before usage. The study compared the performance of manual annotation, crowd coding, numerous dictionaries, and machine learning by using traditional and deep learning algorithms. The study found that the best performance was attained with trained humans. None of the used dictionaries gave acceptable results, and machine learning, profound learning, outperformed dictionary-based methods. However, it does not reach the level of human performance.

More precisely, some studies criticize the performance of automatic annotation of Lexical-based approaches for sentiment analysis. For example, [18] evaluated three general-purpose sentiment analyzers TextBlob, VADER, and Stanford Core NLP Sentiment Analysis. The study used an online health dataset and a general-purpose dataset. The study concluded that none of the used general-purpose sentiment analyzers produced satisfactory classifications. Also, the result of sentiment analyzers was inconsistent when applied to the same dataset. [19] worked with Unsupervised Approach, a lexical approach using open-source libraries such as TextBlob, VADER. The study concluded that the results obtained using a lexical-based approach were not accurate with the social media text. In addition, the dataset classified using in-built libraries needs to be classified using supervised machine learning algorithms to obtain accurate results and acceptable accuracy.

Tab. 1 presents a comparative analysis of sentiment analysis models discussed in the literature review.

Table 1: Summary of sentiment analysis models and techniques discussed in the literature

Reference	Model	Dataset	Result
[3]	TextBlob, VADER, and Stanford Core	News from a total of ten newspapers and five websites that covers a wide variety of economic and financial issues	None of the used general-purpose sentiment analyzers produced satisfactory classifications
[14]	TextBlob	A large-scale dataset, COVIDSENTI that consists of 90 000 tweets that were related to COVID-19	The study concluded that there is a need to develop a proactive and agile public health presence to combat the spread of negative sentiment on social media following a pandemic
[15]	TextBlob and VADER	Patients' opinions dataset	VADER lexicon-based approach outperformed TextBlob model
[16]	NLP and different machine learning classifiers like decision tree, random forest, k-nearest neighbors, and Gaussian naïve Bayes	Dataset of 2500 tweets.	The random forest classifier was the most accurate model
[17]	Machine learning techniques	Student's feedback before and after the COVID-19 pandemic	the average accuracy of 85.62% was achieved by the Support vector machine algorithm
[18]	Manual annotation, crowd coding, numerous dictionaries, and machine learning	Online health dataset and a general-purpose dataset	None of the used dictionaries gave acceptable results, and machine learning, profound learning, outperformed dictionary-based methods
[19]	TextBlob and VADER	Dataset of tweets related to events like Howdy Modi, a gathering in Houston, Haryana assembly elections, and movies releases.	The results obtained using the lexical-based approach were not accurate

This study focused on proving that the automatic annotation using existing lexicons sentiment does not reach the level of human annotation performance.

3 Methodology

This section explains the study's methodology, divided into five stages (data collection, data cleaning, data translation, manual data annotation, automatic data annotation, and classification and evaluation). Also, it covers the techniques that were used in each phase. Fig. 1 shows the used methodology.

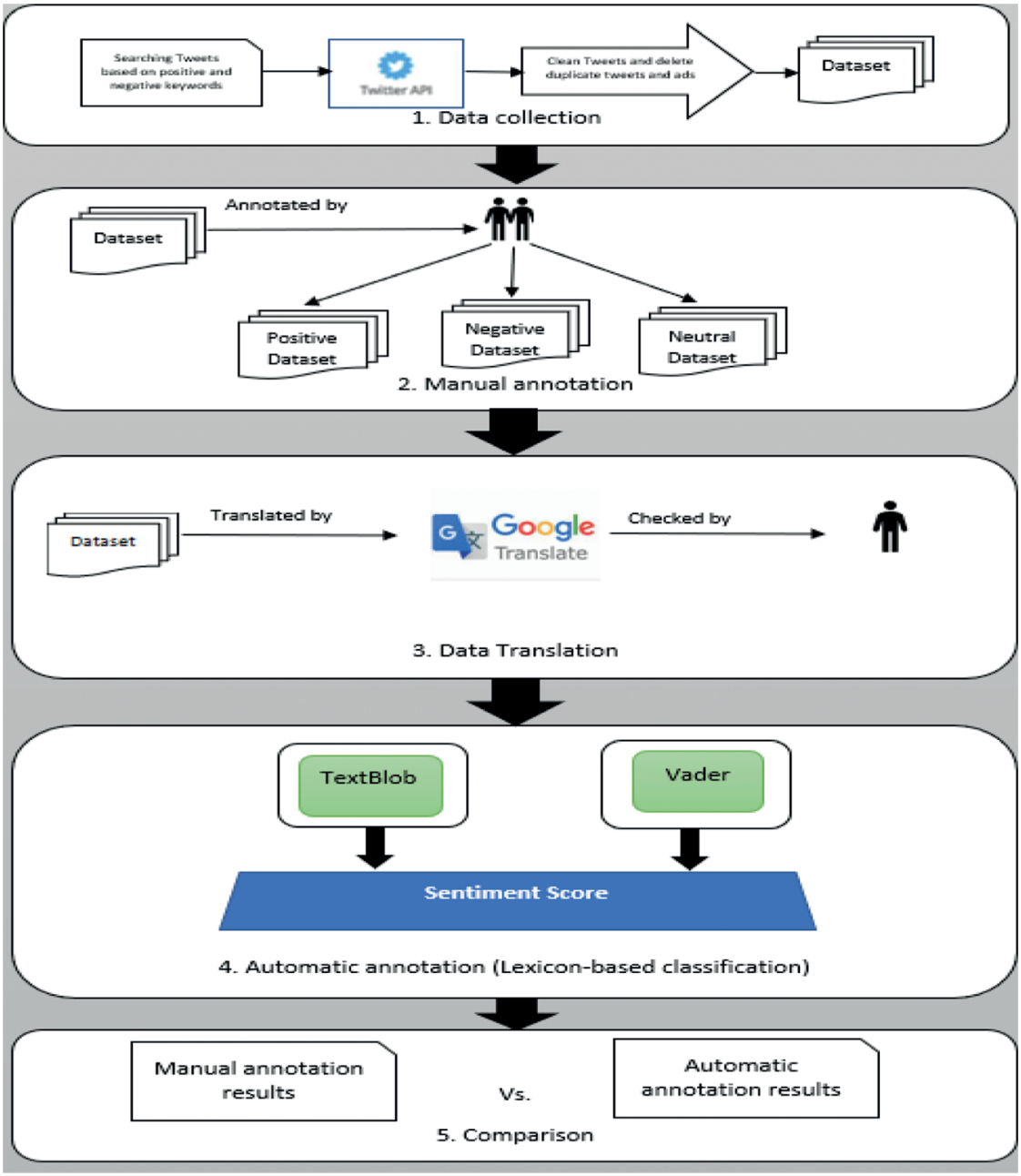

Figure 1: The methodology of the study

3.1 Data Collection

Twitter's full-archive search API collected 25800 Arabic tweets using positive, neutral, and negative keywords. Examples of the positive research keywords are shown in Tab. 2.

Table 2: Positive words

Word	Translation
آمن	Safe
فعال	Effective
جيد	Good
يزيد المناعة	Increase immunity
ممتاز	Excellent

Examples of the negative research keywords are shown in Tab. 3.

Table 3: Negative words

Word	Translation
غيرآمن	Unsafe
غيرفعال	Inefficient
سيئ	Bad
تجلط	Clot
جلطة	Stroke
طفح جلدي	Rash
قاتل	killer
خطير	Dangerous
مميت	Fatal, deadly
سم	Poison

Examples of the neutral research keywords are shown in Tab. 4.

Table 4: Neutral words

Word	Translation
عدد الملقحين	Number of vaccinated
يتكون اللقاح من	vaccine consist of
يبدأ التطعيم	Vaccination starts
اللقاحات المتوفرة	Available vaccines
المعلومات الطبية للقاح	Medical information of vaccine
خطة التطعيم	Vaccination plan
أنواع اللقاح المتوفرة	Type of vaccines

25800 tweets were collected using the keywords mentioned above. However, one delicate and essential point must be mentioned here to avoid confusion and misunderstanding. One may be confused that the positive keywords can indicate positive tweets; however, it is not the case. Sometimes the word is entirely positive such as "safe," but it can be used in negative tweets for sarcasm.

Example of a positive tweet using the "Safe" keyword: "Pfizer is **safe**, God willing, and has no side effects." Example of a negative tweet using the "Safe" keyword: "Good evening, doctor, one of the reasons for not receiving the vaccine from some is the fear of clots, and everyone fears of AstraZeneca, despite the declarations that it is **safe**, but such news frightened people."

3.2 Data Cleaning

In this phase, 25800 raw tweets were cleaned by deleting duplicate tweets. The resulted tweets were read one by one by humans to remove ads and tweets unrelated to the vaccines. After that, each tweet was cleaned by deleting the links and symbols like "#" and "@."

3.3 Manual Data Annotation

The tweets in the dataset were manually annotated into three classes: positive, negative, or neutral, based on the text and the tweet's context. Then the tweet was saved as annotated text in an Excel sheet. Tweets annotation was a complex process. Thus, some rules were used to annotate tweets as positive, negative, or neutral. Some of these rules are as follows:

- The tweets with negative words about the vaccination or any vaccine were annotated as negative.
- The tweets about substantial side effects of the vaccine-like clot brain stroke, heart stroke, or admission to intensive care were annotated as negative.
- The tweets with positive words about the vaccination or any type of vaccine were annotated as positive.
- Tweets mentioning vaccination or any form of a vaccine and tweets about the number of persons who have been vaccinated were labeled as neutral.

3.4 Data Translation

The dataset was translated from Arabic to English because TextBlob and VADER lexicons cannot work on Arabic text. The translation was performed using Google translation. Then, the validity of translation was checked by humans. The tweets with wrong translations were deleted, and those with incomplete translations were modified.

Example of wrong translation:

The tweet:

"كلهم زينات بس احسن شي فايزر"

The wrong translation:

"They are all decorations, but the best thing is Pfizer."

The correct translation:

"All of them are good, but Pfizer is the best."

Example of incomplete translation:

The tweet:

"أنا مؤيد جدا للتطعيم ولكن ضد التطبيل ان هو آمن 100% وضد جماعة نظريات المؤامرة شركة استرازينكا "وهذا امر طبيعي في اي دوا" أعراضهم قوية جدا واخرها الجلطة ولو كانت بنسبة ضعيفة اللي قرر ان يتطعم وعنده اختيار الشركة يفوت استرازينكا، فايزر وسبونتيك اكثر أمانا واقل أعراض"

The translation:

"I am very supportive of vaccination, but against drumming that it is 100% safe and against a group of conspiracy theories, AstraZeneca company. "This is normal in any drug."

All underlined text was not translated. When we spilled the sentence into two sentences, all the sentences were translated.

Modified tweet:

"أنا مؤيد جدا للتطعيم ولكن ضد التطبيل ان هو آمن 100% وضد جماعة نظريات المؤامرة شركة استرازينكا "وهذا امر طبيعي في اي دوا" أعراضهم قوية جدا واخرها الجلطة ولو كانت بنسبة ضعيفة اللي قرر ان يتطعم وعنده اختيار الشركة يفوت استرازينكا، فايزر وسبونتيك اكثر أمانا واقل أعراض"

The translation of modified tweet:

"I am very supportive of vaccination, but against drumming that it is 100% safe and against the conspiracy theorists group AstraZeneca "and this is normal in any drug" their symptoms are very strong, and the last of them is the clot, even if it is a weak percentage, who decided to vaccinate and has the choice of the company, miss AstraZeneca, Pfizer and Sputnik are safer and less symptoms"

3.5 Automatic Data Annotation

This phase automatically classifies the manually annotated tweets to positive, negative, and neutral using TextBlob and VADER sentiment lexicons.

3.5.1 TextBlob Classification

As mentioned in the literature review section, TextBlob is an open-source Python library used for classifying textual data. It uses two measures to analyze the sentiment of the text polarity and subjectivity. Polarity determines if the text expressed positive, negative, or neutral sentiments. It is a number between [−1, 1], −1 indicates negative sentiments, +1 indicates positive sentiment, and 0 indicates neutral sentiment. Subjectivity determines if the tweet relies on facts, opinions, assumptions, or influences by personal feelings. It is a number in the range of [0, 1] [5].

Example of TextBlob scoring:

The tweet:

"وكالة الأدوية الأوروبية تعطي الضوء الأخضر للقاح أسترازينيكا: آمن وفعّال"

"The European Medicines Agency gives the green light to the AstraZeneca vaccine: safe and effective."

Polarity: 0.26

Subjectivity: 0.46

3.5.2 VADER Classification

The valence aware dictionary and sentiment reasoned (VADER) is a python package used to analyze text. First, the Sentiment Intensity Analyzer was loaded from the VADER package. Then the polarity scores method was used to get the sentiment scores (positive, negative, neutral, and compound scores) of the tweets [20]. It considers the context of the tweets and how the words are written. The compound score is the metric used to give the tweet's sentiment; it is a number that ranges from −1 to 1. The sentiment is positive if compound greater than or equal 0.05, neutral if compound between [−0.05, 0.05], and negative if compound less than or equal −0.05 [20].

Example of VADER scoring:

The tweet:

"وكالة الأدوية الأوروبية تعطي الضوء الأخضر للقاح أسترازينيكا: آمن وفعّال"

"The European Medicines Agency gives the green light to the AstraZeneca vaccine: safe and effective."

VADER scored this sentence as: {'neg': 0.0, 'neu': 0.684, 'pos': 0.316, 'compound': 0.7184}

1) Capitalization increases the intensity of positive or negative scores.

"The European Medicines Agency gives the green light to the AstraZeneca vaccine: SAFE and EFFECT."

VADER scored this sentence as: {'neg': 0.0, 'neu': 0.635, 'pos': 0.365, 'compound': 0.8159}

2) Exclamation marks increase the intensity of sentiment scores.

"The European Medicines Agency gives the green light to the AstraZeneca vaccine: SAFE and EFFECT!"

VADER scored this sentence as: {'neg': 0.0, 'neu': 0.626, 'pos': 0.374, 'compound': 0.8298}

3) The words present before the positive or the negative word increase or decrease the intensity of the sentiment.

"The European Medicines Agency gives the green light to the AstraZeneca vaccine: SAFE and EFFECT!"

VADER scored this sentence as: {'neg': 0.0, 'neu': 0.627, 'pos': 0.373, 'compound': 0.8524}

4) If the text contains 'but,' the sentiments before and after 'but' are considered; however, the sentiment after "but" is weighted more heavily than before "but."

"The European Medicines Agency gives the green light to the AstraZeneca vaccine: SAFE but EFFECT!"

VADER scored this sentence as: {'neg': 0.0, 'neu': 0.625, 'pos': 0.375, 'compound': 0.8555}.

3.6 Manual Annotation and Automatic Annotation Results Comparison

The results of automatic annotation using TextBlob, and VADER were compared to the manual annotation to check the performance of the algorithms.

3.7 Classification Matrix

A classification matrix is a tool used to assess the results of prediction. It sorts all cases from the classifier into categories (true positive, false positive, true negative, false negative by evaluating whether the predicted values matched the actual values. Then the total of all cases in each category is displayed in a matrix. It is a standard tool for evaluating the valuation of models and is referred to as a *confusion matrix*. Fig. 2 shows the classification matrix categories [21].

- True positives (TP): The cases in which the tweets were predicted as positive and, they are positive
- True negatives (TN): The cases in which the tweets were predicted as negative while negative.
- False positives (FP): The cases in which the tweets were predicted as positive while negative.
- False negatives (FN): The cases in which the tweets were predicted as negative and, they are positive.

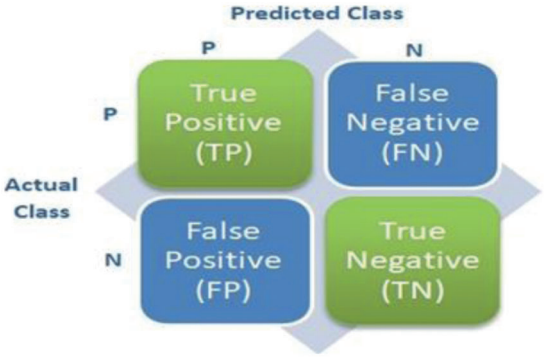

Figure 2: Classification matrix

Precision (positive predictive value): is the percentage of tweets that were identified positive are really positive. Eq. (1) used to calculate precision:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

Recall (sensitivity) is the percentage of correctly predicted tweets from all the positive classes. Eq. (2) calculated recall:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

The accuracy is the overall success rate [21]. From all the positive and negative classes, how many of them we have predicted correctly. Eq. (3) calculated precision:

$$\text{Recall} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3}$$

F-Measure is a harmonic average of precision and recall value [22]. It calculates the overall performance of a classifier, and it is calculated by Eq. (4).

$$\text{F-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

ROC stands for "Receiver operator characteristic curve," a graph that shows the performance of a classification model at all classification thresholds. It plots two parameters true positive rate (TPR) and false positive rate (FPR) [23]. TPR was measured using Eq. (5), and FPR was measured using Eq. (6).

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{6}$$

AUC stands for "Area under the ROC Curve." It provides an aggregate measure of performance across all possible classification threshold. Hosmer and Lemeshow [22] suggest AUC of:

50%-60% = Fail.

60%-70% = Poor.

70%-80% = Fair.

80%-90% = Good.

90%-100% = Excellent.

4 Results: Analysis and Discussion

The results will be presented in two sections. Section 4.1 will present the result of sentiment library TextBlob on the datasetwith a discussion of the classification performance. While Section 4.2 will show and discuss the result of sentiment library VADER on the dataset.

4.1 TextBlob Sentiment Results and Analysis

4.1.1 Objective

The objective is to get the sentiment scores of the manually annotated tweets using the TextBlob library and examine the accuracy and performance of the sentiment classification.

4.1.2 Method

The dataset, after all, preprocessing steps, consists of approximately 5402 tweets. There were 3124 tweets annotated as positive, 1463 annotated as negative, and 815 tweets annotated as neutral. TextBlob classification included many steps as follows:

- The sentiment score of each tweet was calculated by measuring the sentiment of each word in the tweet.
- The tweets were annotated as positive, negative, and neutral based on their sentiment scores. For TextBlob If the score was less than zero; the tweet was annotated as negative. If the tweet was greater than zero; the tweet was annotated as positive. If the tweet was equal to zero; the tweet was annotated as neutral.
- For evaluating, TextBlob classification results (polarity scores) were used to discuss the classification performance.

4.1.3 Results

The following Tab. 5 and Fig. 3 indicate applying the TextBlob library on the dataset, including positive, negative, and neutral tweets.

Table 5: TextBlob results

TextBlob–(positive class)							
		Confusion matrix		Result			
		pos	neg	Precision	Recall	F-Measure	Accuracy
Original dataset	pos	2831	293	0.68%	0.91%	0.78%	0.70%
	neg	1330	948				
TextBlob–(negative class)							
		Confusion matrix		Result			
		pos	neg	Precision	Recall	F-Measure	Accuracy
Original dataset	pos	339	1124	0.54%	0.23%	0.32%	0.74%
	neg	287	3652				

(Continued)

Table 5 (continued)

TextBlob–(neutral class)							
		Confusion matrix		Result			
		pos	neg	Precision	Recall	F-Measure	Accuracy
Original dataset	pos	269	546	0.44%	0.33%	0.38%	0.83%
	neg	346	4241				

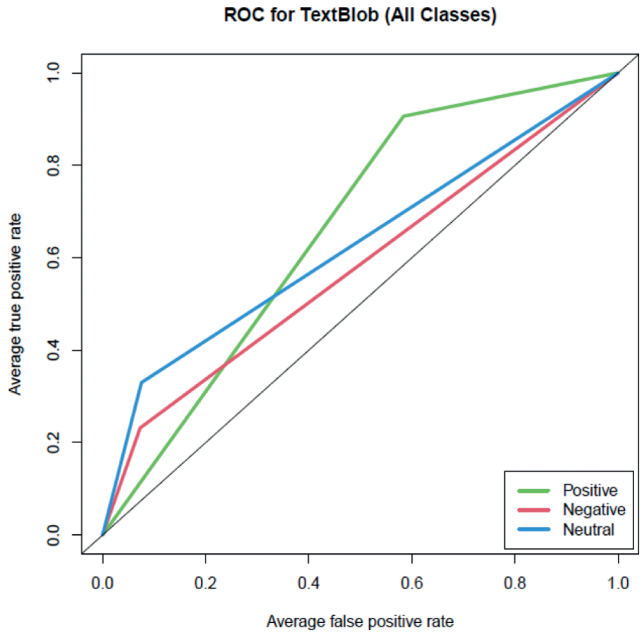

Figure 3: ROC of TextBlob for the three classes positive, negative, and neutral

4.1.4 Discussion

The results of the TextBlob performance are shown in the previous section using Tab. 5 and Fig. 3 Based on Tab. 5:

For the positive class, the classification was successful for 2831 cases from 3124 (tweets manually annotated as positive). It was failed for 293 cases from 3124 (tweets manually annotated as negative). For the negative class, the classification was successful for 339 cases from 1463 (tweets manually annotated as positive), and it was failed for 287 cases from 1463 (tweets manually annotated as negative). For the neutral class, the classification was successful for 269 cases from 815 (tweets manually annotated as positive). It was failed for 346 cases from 815 (tweets manually annotated as negative). Successful classification means the classification agrees with the actual manual labels, and failed classification does not agree with the actual labels.

Based on Fig. 3 ROC curve of the positive class was better than the negative and neutral class. Further, Tab. 6 indicated that the AUC of positive and neutral classes was poor. In contrast, the AUC of the negative class was failed, which indicated that the AUC of all classes was not acceptable.

Table 6: TextBlob AUC

Class	AUC	Hosmer and lemeshow suggestion
Positive	66.12%	Poor
Negative	57.94%	Fail
Neutral	62.73%	Poor

4.2 VADER Sentiment Results and Analysis

4.2.1 Objective

The objective is to get the sentiment scores of the dataset using the VADER sentiment library.

4.2.2 Method

The used dataset consists of approximately 5402 tweets. There were 3124 tweets annotated as positive, 1463 annotated as negative, and 815 tweets annotated as neutral. VADER classification included many steps as follows:

- The sentiment score of each tweet was measured by calculating the sentiment score of each word in the tweet.
- The tweets were annotated as positive, negative, and neutral based on their sentiment scores. For VADER, the tweets were annotated as positive if the compound score was greater than or equal to 0.05 and negative if the compound score was less than or equal to −0.05. While the tweets were annotated as neutral if the compound score was between [−0.05, 0.05].
- For evaluating, the sentiment score (compound) was used to get the classification accuracy.

4.2.3 Results

Tab. 7 and Fig. 4 indicate the results of applying the VADER lexicon to the dataset.

Table 7: VADER results

VADER–positive class							
		Confusion matrix		Result			
		pos	neg	Precision	Recall	F-Measure	Accuracy
Original dataset	pos	2598	526	0.68%	0.83%	0.75%	0.68%
	neg	1216	1062				
VADER-negative class							
		Confusion matrix		Result			
		pos	neg	Precision	Recall	F-Measure	Accuracy
Original dataset	pos	700	763	0.48%	0.57%	0.52%	0.76%
	neg	529	3410				
VADER-neutral class							
		Confusion matrix		Result			
		pos	neg	Precision	Recall	F-Measure	Accuracy
Original dataset	pos	85	730	0.24%	0.10%	0.14%	0.81%
	neg	274	4313				

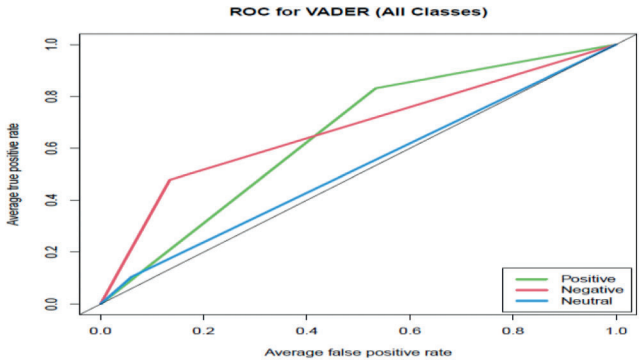

Figure 4: ROC of VADER for the three classes positive, negative, and neutral

4.2.4 Discussion

The results of VADER performance are detailed in the previous section using Tab. 6 and Fig. 4. Based on Tab. 6, for the positive class, the classification was successful for 2598 cases from 3124 (tweets manually annotated as positive). It was failed for 1216 cases from 3124 (tweets manually annotated as negative). For the negative class, the classification was successful for 700 cases from 1463 (tweets manually annotated as positive), and it was failed for 529 cases from 1463 (tweets manually annotated as negative). For the neutral class, the classification was successful for 85 cases from 815 (tweets manually annotated as positive). It was failed for 274 cases from 815 (tweets manually annotated as negative). As mentioned in Section 4.1.4, successful classification means the classification agrees with the actual manual labels and failed classification does not agree with the true labels.

Based on Fig. 4 ROC curve of the positive class was better than the negative and neutral class. Further, Tab. 8 indicated that the AUC of positive and negative classes was poor based on Hosmer and Lemeshow. The neutral class was failed, which means the AUC of all classes was not acceptable.

Table 8: VADER AUC

Class	AUC	Hosmer and lemeshow suggestion
Positive	64.89%	Poor
Negative	67.21%	Poor
Neutral	52.23%	Fail

4.3 Automatic Annotation vs. Manual Annotation

Both TextBlob and VADER classifications better predict positive tweets than negative and neutral ones. For positive class, TextBlob accuracy by 70% outperforms VADER accuracy by 68%. VADER accuracy by 76% for the negative class outperforms TextBlob accuracy by 74%. TextBlob accuracy by 83% outperforms VADER accuracy by 81% for the neutral class.

In TextBlob for positive class, the classification was unsuccessful for 293 cases from 3124. In contrast, VADER's classification was unsuccessful for 526 cases from 3124. Some of the examples in the positive class where TextBlob and VADER classifications were not successful are:

Example 1:

The tweet: "Jordan-Jordanian specialists: AstraZeneca vaccine is safe: (MENAFN-Khaberni) Khaberni-Specialists in the field of vaccines and epidemiology said that there is no scientifically proven reason for fear and hesitation from using the AstraZeneca anti-virus vaccine."

TextBlob polarity: −0.125

VADER compound score: −0.5574

The tweet was annotated by as positive.

Example 2:

The tweet: "The European Union, Britain, and America indicated the effectiveness of the vaccine and its benefits that exceed any risks within a small percentage that was infected."

TextBlob polarity: −0.125

VADER compound score: −0.4019

The tweet was annotated by human annotated it as positive.

In TextBlob for the negative class, the classification was unsuccessful for 1124 cases from 1463. In VADER, it was not successful for 763 cases from 1463. Some of the tweets in the negative class where TextBlob and VADER classifications were not successful are:

Example 1:

The tweet: "Even Pfizer causes clots, my mother is now sleeping in the care after a stroke after he took the first dose of the Pfizer vaccine."

TextBlob polarity: 0.25

VADER compound score: 0.4939

The tweet was annotated by TextBlob and VADER as positive while human-annotated it as negative.

Example 2:

The tweet: "Monday April 19 Bulgaria: AstraZeneca stopped giving women under the age of 60, as they have an increased risk of stroke Sweden: A prominent doctor stated that hundreds of AstraZeneca doses are disposed of daily in Stockholm, because no one wants to get them AstraZeneca Vaccine. Saudi Arabia: 15 cases of blood clots from the AstraZeneca vaccine"

TextBlob polarity: 0.233

VADER compound score: 0.1531

The tweet was annotated by TextBlob and VADER as positive, while human-annotated it as negative.

In TextBlob for neutral class, the classification was unsuccessful for 546 cases from 815. In comparison, it was not successful in VADER for 730 cases from 815. Some of the tweets in the neutral class where TextBlob and VADER classifications were not successful are:

Example 1:

The tweet: "The number of people vaccinated with a dose is starting to drop because they have taken the second dose. I mean, they moved from the dose phase and became in the statistics of the second dose. The numbers are correct."

TextBlob polarity: −0.078125

VADER compound score: −0.2023

The tweet was annotated by both TextBlob and VADER as negative while it was annotated by human as neutral.

Example 2:

The Tweet: "I imagine the numbers are wrong, the number of vaccinated people in Iraq is approximately one million and 700 thousand, and the population of Iraq is approximately 41 million, meaning the vaccination rate is more than 4 percent."

TextBlob polarity: −0.2

VADER compound score: −0.4215

The tweet was annotated by as neutral.

The previous examples of TextBlob and VADER classifications failure led to some drawbacks of TextBlob and VADER algorithms. The drawbacks are:

1) If the word was not found in the sentiment lexicon, it was neutral.

Example 1:

The tweet: "Stroke in the eye from Pfizer vaccination" is a negative word if used with the COVID-19 vaccine.

VADER is considered 'neu' with a 0.0 compound score. Furthermore, TextBlob is considered 'neu' with a polarity score equal to 0.0. the two algorithms did not consider "stroke" a negative word.

Example 2:

The tweet: "I took the second dose of Pfizer on March 29, and now I have a stomachache."

"stomachache" is considered a negative word because it is a type of pain. The compound score in VADER equals 0.0, while the polarity in TextBlob equals 0.0.

2) The main drawback of the rule-based approach (TextBlob and VADER) for sentiment analysis is that the method cares about individual words but ignores the context in which it is used.

Example 1:

The tweet: "Hello, I'm Pfizer, is your immune system bad? Instead of improving and strengthening it, do not worry, we have the solution to Corona disease, because our vaccines have proven to be safe and effective, although they are still in the middle of clinical trials. Oh, did we not tell you that you are under trial, and we are not responsible for any damage or death that you suffer due to our genetic treatment!"

In TextBlob, the polarity score of the whole tweet equals 0.0239, which is annotated as 'neu', while in VADER, the compound score equals 0.2815, which is annotated as 'pos.'

Part of the tweet is the sentence "Oh did we not tell you that you are under trial and we are not responsible for any damage or death that you suffer due to our genetic treatment!"

The polarity score of TextBlob equals −0.128125, which is annotated as 'neg', while VADER compound score equals −0.9162, which also is annotated as neg.

Example 2:

The tweet: "According to the magazine, the "Pfizer" vaccine that contains these proteins has taken full and emergency approval in the United States of America without serious testing for the safety of the product on humans during the coming years. And the worst wave of deaths was recorded among those who received the "Pfizer" vaccine in Norway."

In TextBlob, the polarity score of the tweet equals −0.328, which is annotated as 'neg' and in VADER, the compound score equals 0.3009, which is annotated as 'pos'.

Part of the tweet is the sentence, "the worst wave of deaths was recorded among those who received the "Pfizer" vaccine in Norway".

In TextBlob, the polarity score equals −1.0 that is like the annotation of the whole tweet. While in VADER, the compound score equals −0.6249, which is different from the annotation of the whole tweet, which is annotated as positive.

Example 3:

The tweet: "No no no no no no no no no to Pfizer vaccines, and our director, AztraNika. The vaccine is more dangerous than the virus and causes sterility. This is the plan because they said that many people in the world. They want to lack people in the world, especially the Arabs. Understand, Muslims, vaccines are poison in the human body."

In VADER, the compound score of the tweet equals −0.974, while in TextBlob, the polarity score equals 0.0229

Part of the tweet is the sentence, "The vaccine is more dangerous than the virus and causes sterility." This sentence in VADER has a compound score equal to −0.526, while in TextBlob, its polarity score is equal to −0.05. The tweet was annotated as negative by both VADER and TextBlob libraries.

The tweet is, "vaccines are poison in the human body." Its compound score in VADER equal −0.5423, which is like the annotation of the full tweet. Nevertheless, its polarity score in TextBlob equals 0.0, which is annotated as neutral.

3) TextBlob and VADER algorithms are used with the English language so that the Arabic tweets need the translation before classification. One of the big problems faced in Arabic translation is dialects and diacritical marks. If the word is translated wrong, the sentiment scores would be wrong.

Example 1:

The tweet:

"الي معترض علي اخذ لقاحين مختلفين مادري ليش مو عاجبكم

اوكسفورد امن امن امن بشهاده كل مندوبي اللقاحات

فايزر امن امن امن بشهاده كل مندوبي اللقاحات

جرعتكم الثالثه جونسون اند جونسون امن امن امن بشهاده كل مندوبي اللقاحات"

Translation:

"Whoever objects to taking two different vaccines, I do not know why you do not like it

Oxford Security Safe Security With the testimony of all vaccine delegates

Pfizer, security, security, security, according to the testimony of all vaccine representatives

Your third dose, Johnson & Johnson, safe, secure, safe, with the testimony of all vaccine representatives."

The word "امن" has many translations like security, safety, and secure in the previous tweet.

Example 2:

The tweet: "انا احسن منكم. انا ماخذه فايزر وجرعتين. وشايفه نفسي وبتعنصر عليكم 😂"

Translation: "I am better than you. I take Pfizer and two doses. And I see myself, and I hate you 😂"

The word "اتعنصر" translated as "hate" which has a sentiment score equal −0.5 in VADER and-.08 in TextBlob. The correct translation is "racist," with a sentiment score equal to −0.6124 in VADER and 0.0 in TextBlob.

4) TextBlob is not able to detect the negation in all cases.

Example:

The tweet: "pFizer is not the best"

TextBlob polarity= 1.0

VADER compound score= −0.5216

Modified tweet: "pFizer is not best"

TextBlob polarity= −0.5

VADER compound score= −0.5216

VADER algorithms give the sentence the same score, while TextBlob could not detect the negation in the first case.

5) TextBlob algorithms can measure the polarity for adjectives (like safe, lazy, dangerous, etc.). However, they cannot determine the polarity of comparative and superlative adjectives or assign less positivity or negativity to them (such as safer, safest, laziest, dangerous, dangerous, more hazardous), as indicated in Tab. 9.

Table 9: TextBlob polarity for adjectives

Word	TextBlob polarity	VADER compound score
Safe	0.5	0.44
Safer	0.0	0.42
Safest	0.0	0.40
Lazy	−0.25	−0.36
Lazier	0.0	−0.51
Laziest	0.0	−0.57
Dangerous	−0.6	−0.48
More dangerous	−0.05	−0.53
Most dangerous	0.05	−0.53

Example:

The tweet: "The safest vaccine without complications."

TextBlob polarity: 0.0.

VADER compound: 0.4.

Based on its context, the tweet is positive for the Pfizer vaccine, so it was annotated manually as positive. Nevertheless, it was annotated as neutral by TextBlob and VADER. If the word safest changes to safe, the TextBlob polarity will be 0.5, and the VADER compound score will 0.44. So, the automatic annotation will be successful using the VADER algorithm.

5 Conclusion and Future Work

There are different algorithms for text classification and sentiment analysis. This study demonstrates TextBlob and VADER's poor performance for automatic sentiment annotation/classification quantitatively. For positive class, the classification was victorious by 70% using TextBlob and by 75% using VADER. While for the negative class, the classification was victorious by 74% using TextBlob and 76% using VADER. For neutral class, the classification was victorious by 83% using TextBlob and by 81% using VADER.

The ROC curve and AUC results showed that classification models TextBlob and VADER are not reliable for automatic sentiment annotation.

This study discussed many drawbacks and limitations of automatic annotation using lexicon-based algorithms like:

1) The algorithms ignore the words they know and classify them as neutral. They assigned them wrong polarity and averages to get the sentiment score that may not reflect the actual sentiment of the text.
2) TextBlob and VADER ignore the context in which the word is used, which may change the word's sentiment.
3) The sentiment scores of the algorithms rely on the average of the sentiment score of individual words, which may not reflect the actual sentiment of the text.
4) TextBlob and VADER algorithms can be used with the English language. Therefore the Arabic tweets were needed to be translated into English before classification. One of the big problems faced while translating Arabic into English is dialects and diacritical marks. If the word is translated wrong, the sentiment scores would be wrong.
5) Some automatic annotation algorithms cannot detect the negation in all cases to not give an actual sentiment score.
6) Some automatic annotation algorithms can detect polarity in adjectives (such as safe, dangerous, and so on). However, they cannot detect it in comparative and superlative adjectives or assign them less positivity or negative (safer, safer, dangerous, more dangerous).

In the future, we will consider using machine learning methods and deep learning models to check the performance of automatic annotation compared to manual annotation.

Acknowledgement: The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia, for funding this research work through project number 959.

Funding Statement: This work was supported by Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia, for funding this research work through Project Number 959.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] M. Petrillo and J. Baycroft, "Introduction to manual annotation," *Fairview Research*, pp. 1–7, 2010.

[2] G. Olague, M. Olague, A. R. Jacobo-Lopez and G. Ibarra-Vazquez, "Less is more: Pursuing the visual turing test with the kuleshov effect," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 1553–1561, 2021.

[3] W. Atteveldt, M. A. Velden and M. Boukes, "The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms," *Communication Methods and Measures*, vol. 15, no. 2, pp. 121–140, 2021.

[4] S. Vashishtha and S. Susan, "Fuzzy rule-based unsupervised sentiment analysis from social media posts," *Expert Systems with Applications*, vol. 138, pp. 112834, 2019.

[5] V. Bonta and N. K. N. Janardhan, "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.

[6] E. T. K. Sang and A. Bosch, "Dealing with big data: The case of twitter," *Computational Linguistics in the Netherlands*, vol. 3, pp. 121–134, 2013.

[7] A. Mauro, M. Greco and M. Grimaldi, "A formal definition of big data based on its essential features," *Library Review*, vol. 65, no. 3, pp. 122–135, 2016.

[8] A. L'Heureux, K. Grolinger, H. F. Elyamany and M. A. M. Capretz, "Machine learning with big data: Challenges and approaches," *IEEE Access*, vol. 5, pp. 7776–7797, 2017.

[9] U. A. Siddiqua, T. Ahsan and A. N. Chy, "Combining a rule-based classifier with ensemble of feature sets and machine learning techniques for sentiment analysis on microblog," in *19th Int. Conf. on Computer and Information Technology (ICCIT)*, USA, 2016.

[10] B. Awrahman and B. Alatas, "Sentiment analysis and opinion mining within social networks using konstanz information miner," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 9, no. 1, pp. 15–22, 2017.

[11] M. A. Musen, C. A. Bean, K. H. Cheung, M. Dumontier, K. A. Durante *et al.,* "The center for expanded data annotation and retrieval," *Journal of the American Medical Informatics Association*, vol. 22, no. 6, pp. 1148–1152, 2015.

[12] M. Neves and J. Ševa, "An extensive review of tools for manual annotation of documents," *Briefings in Bioinformatics*, vol. 22, no. 1, pp. 146–163, 2021.

[13] J. C. Lyu, E. Han and G. K. Luli, "Covid-19 vaccine–related discussion on twitter: Topic modeling and sentiment analysis," *Journal of Medical Internet Research*, vol. 23, no. 6, pp. e24435, 2021.

[14] U. Naseem, I. Razzak, M. Khushi, P. W. Eklund and J. Kim, "Covidsenti: A large-scale benchmark twitter data set for covid-19 sentiment analysis," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 1003–1015, Aug. 2021.

[15] V. I. S. RamyaSri, C. Niharika, K. Maneesh and M. Ismail, "Sentiment analysis of patients & opinions in healthcare using lexicon-based method," *International Journal of Engineering and Advanced Technology*, India, vol. 9, no. 1, pp. 6977–6981, 2019.

[16] M. Wadera, M. Mathur and D. K. Vishwakarma, "Sentiment analysis of tweets-a comparison of classifiers on live stream of twitter," in *4th Int. Conf. on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, pp. 968–972, 2020.

[17] M. Umair, A. Hakim, A. Hussain and S. Naseem, "Sentiment analysis of students' feedback before and after covid-19 pandemic," *International Journal on Emerging Technologies*, vol. 12, no. 2, pp. 177–182, 2021.

[18] L. He and K. Zheng, "How do general-purpose sentiment analyzers perform when applied to health-related online social media data?," *Studies in Health Technology and Informatics*, vol. 264, pp. 1208, 2019.

[19] S. Zahoor and R. Rohilla, "Twitter sentiment analysis using lexical or rule based approach: A case study," in *8th Int. Conf. on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, India, pp. 537–542, 2020.

[20] C. H. E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *8th Int. Conf. on Weblogs and Social Media (ICWSM-14)*, Ann Arbor, Michigan, USA, 2014, Available at: http://comp.social.gatech.edu/papers/icwsm14.

[21] J. Brownlee, "How to calculate precision, recall, and F-measure for imbalanced classification," Available: https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalancedclassification, 2020.

[22] P. A. Flach and M. Kull, "Precision-recall-gain curves: Pr analysis done right," in *NIPS*, Bristol, United Kingdom, vol. 15, 2015.

[23] S. Menard, "Applied logistic regression analysis," *Sage*, vol. 106, pp. 88, 2002.