Hybrid Approach with VADER and Multinomial Logistic Regression for Multiclass Sentiment Analysis in Online Customer Review

Murahartawaty Arief, Noor Azah Samsudin
Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia,
Batu Pahat, Johor, Malaysia

***Abstract*—Sentiment analysis is crucial for businesses to understand customer reviews and assess sentiment polarity. A hybrid technique combining VADER and Multinomial Logistic Regression was used to analyze customer sentiment in online customer review data. VADER is a lexicon-based approach that labels reviews with sentiment using a predefined lexicon, whereas Multinomial Logistic Regression can determine the polarity of sentiment using VADER data. This study employed multiclass classification using TF-IDF vectorization to categorize sentiment as a positive, negative, or neutral class. Correctly managing neutral sentiments can assist businesses in identifying improvement opportunities. The utilization of the VADER lexicon and Multinomial Logistic Regression has been shown to significantly improve the performance of sentiment analysis in the context of multiclass classification problems. With a 75.213% accuracy rate, the VADER lexicon accurately recognizes neutral sentiment and is appropriate to adapt in categorizing sentiment related to customer reviews. Combined with Multinomial Logistic Regression, accuracy increases to 92.778%. In conclusion, the hybrid approach with VADER and Multinomial Logistic Regression can leverage the accuracy and reliability of multiclass customer sentiment analysis.**

***Keywords*—Hybrid approach; multiclass sentiment analysis; VADER; multinomial logistic regression; online customer review**

I. INTRODUCTION

Understanding and analyzing customer sentiment become a demanding topic in the sentiment analysis research area. Sentiment analysis is crucial for businesses to detect and extract customer reviews to determine consumer sentiment and measure satisfaction levels based on the sentiment expressed in the reviews. Sentiment Analysis uses statistical approaches, natural language processing, and machine learning to analyze and classify customer sentiment as positive, negative, or neutral [1].

Machine learning has made significant progress in classifying customer sentiment in online reviews. Unfortunately, the main focus of this study has concentrated on determining binary classification: positive and negative sentiment. The neutral sentiment is frequently ignored or removed. In an online review, a neutral class is derived from a 3-star rating for a customer who is satisfied enough with the quality of the product. For example: *"Would like to see better battery life. Decent otherwise."* The word "decent" lacks strong positive or negative sentiment about the product being reviewed. This review categorizes it as a neutral class with a mixture of both negative and positive feedback. Neutral class in customer sentiment analysis that allows comments data with factual, informative, or descriptive text rather than expressing a particular emotional tone.

Neutral sentiments should be handled or treated correctly to leverage the comments left and pinpoint areas of improvement. Many businesses pay close attention to reviews on both ends of the scale (4, 5 stars rating as positive and 1, 2 stars as negative) but must catch up on the valuable middle parts (3 stars as neutral). The number of 3 stars in online reviews might be small in volume but higher in value. According to Al-Rubaiee et al. [2], the neutral class can be helpful in sentiment analysis to maintain the accuracy of sentiment analysis models by reducing the risk of misclassifying text that does not express a clear sentiment.

From the business's viewpoint, neutral sentiment should be managed or addressed correctly to acquire a complete understanding of consumer feedback and find areas for improvement to make the product or services stand out. Businesses must guarantee that neutral sentiments are appropriately fulfilled to avoid consumer dissatisfaction, meet baseline requirements, and make well-informed decisions to encourage business growth.

The complexity of natural languages and the difficulty of quantifying human feelings make sentiment analysis a challenging task to categorize text data into different emotion classes automatically. Currently, research on sentiment classification is dominated by two basic approaches: machine learning and lexicon-based approaches. The machine learning approach requires a lot of labeled data training with manual process annotation, reviewing, and verification, which can slow system development and deployment.

Manual labeling using rating values were 1, 2 stars as negative sentiment, 3 stars as neutral sentiment, and 4-5 stars as positive sentiment is not possible to label the review sentences because customers often give a rating that does not match the review. For example, if the experience is valued with 5 stars (excellent), the sentence review has a negative connotation and vice versa. Some customer has their own biases when writing a review and these biases in rating may result in inconsistent reviews. Hu et al. [3] also claimed that customers are not entirely rational, and that affects self-selection rating biases.

The number of customer reviews for a popular product can be hundreds or thousands. Manual labeling commonly used in sentiment analysis is considered inefficient in terms of time and cost, especially if the data is extensive. Conversely, the lexicon-based approach can deal with language complexity and focuses on words and phrases as indicators of semantic orientation. However, this approach is low in accuracy but is computationally efficient, scalable, and provides consistent performance [4]. Combining lexicon-based and machine-learning techniques, known as the hybrid approach, can enhance performance accuracy in sentiment analysis results [5].

Multiple machine learning and lexicon-based approaches have been used to perform automatic classification for sentiment analysis. Unfortunately, the relative effectiveness of each approach is still unclear. This study employed a hybrid approach integrating the VADER lexicon with Multinomial Logistic Regression to examine the sentiment polarity of online customer reviews. The VADER methodology was selected as a lexicon-based approach that uses labels assigned in customer reviews for sentiment classification, utilizing a pre-defined vocabulary. Combined with Multinomial Logistic Regression as a supervised machine learning classifier, it uses a labeled sentiment from the VADER lexicon to train classifiers to generate more accurate sentiment predictions. According to Ramya and Rao [6], Multinomial Logistic Regression can make predictions in big data sets containing various diverse domains.

This research contributes to an extensive experiment with VADER and Multinomial Logistic Regression to determine how this combination affects the performance accuracy to detect neutral classes in sentiment classification. This process used the Amazon product review dataset to ensure the proposed approach is functional and executable in the customer sentiment domain. The performances, advantages, and limitations of VADER with Multinomial Logistic Regression in multiclass sentiment classification tasks were investigated and evaluated in this study. The TF-IDF method was selected as features vectorization to determine the important words to predict the sentiment in the Multinomial Logistic Regression algorithm.

Details about the relevant theories and related works are presented in Section II. Section III presents the research methodology for implementing a hybrid sentiment analysis approach. Experiment, results, and discussion are provided in section IV. Finally, Section V concludes with a conclusion and future works.

II. Related Works

A. Lexicon-based Approach with VADER

Sentiment classification uses automatic algorithms to predict the sentiment orientation of opinions included within a written document, such as a product review, blog post, or social media comment. Sentiment orientation can be characterized as positive, negative, or neutral, or it can be scored on a scale. The lexicon-based approach, also known as the dictionary-based approach, is employed in Natural Language Processing (NLP) for sentiment detection and text classification. This approach involves utilizing a predetermined lexicon or dictionary to identify specific words within the text [7]. A dictionary is prepared in advance, including entries for words or phrases linked to specific categories or sentiments. The lexicon can be established manually by domain experts or generated using automated approaches like machine learning. A score or label reflecting sentiment or category is given to each word or phrase in the lexicon. By aggregating the scores or labels of all the words or phrases in the text, an overall sentiment or category can be determined based on the number and intensity of positive and negative words encountered.

In this research, Valence Aware Dictionary and sEntiment Reasoner (VADER) is used in a lexicon-based approach. VADER is a lexicon-based algorithm developed by Hutto and Gilbert [8] in 2014 to solve the problem of analyzing language, symbols, and style of texts in sentiment analysis. It is widely used in various applications such as social media monitoring, customer feedback analysis, and brand reputation management. It utilizes a pre-built lexicon that contains words with sentiment scores and incorporates grammar and syntactical rules to handle negations, intensifiers, and modifiers [9].

VADER utilizes a pre-built lexicon that contains words or phrases with sentiment scores ranging from -1 (extremely negative) to +1 (extremely positive). The lexicon also includes words with neutral sentiment scores. The sentiment scores in VADER are based on human-annotated ratings and consider the intensity of sentiment associated with each word. The main output of VADER is a sentiment polarity score, which represents the overall sentiment expressed in each text. The score is a continuous value ranging from -1 to +1, where negative values indicate negative sentiment, positive values indicate positive sentiment and values close to zero indicate neutral sentiment [8].

Shabi [10] evaluates the performance of five lexicons used in sentiment analysis on Twitter data: VADER, SentiWordNet, SentiStrength, Liu and Hu opinion lexicon, and AFINN-111. By using the Stanford dataset, the results showed that the best performance in terms of accuracy was achieved by the VADER lexicon 72%, while the performance accuracy of the SentiStrength (67%), AFINN-111 (65%), Liu-Hu lexicon (65%), and SentiWordNet lexicon fell to the value 53%. With this comparison, the lexicon VADER has a good possibility for classifying short text pre-processing and can deal with multi-class, such as positive, negative, and neutral.

Furthermore, Heaton [11] conducts a comparative analysis of TextBlob and VADER in terms of sentiment analysis for positive, negative, and neutral sentiments expressed in social media regarding the NHS Covid-19 applications. The findings indicate that the VADER method outperforms recognized positive sentiments in sarcastic tweets with values of 0.8316, 0.7622, 0.6767, and 0.6958 for four tweets.

B. Machine Learning Approach with Multinomial Logistic Regression

Machine learning is an approach that creates models and designs algorithms to facilitate computational learning and

decision-making processes. Machine learning has progressed beyond teaching computers to imitate the human brain to discovering statistical patterns in learning processes to deliver insights from datasets [12]. Machine learning can be classified into two categories: supervised and unsupervised learning. Supervised learning involves training a model using labelled data, where corresponding target labels or outcomes match the input data. These models are evaluated based on predictive capacity and variance measures. The objective is to acquire the knowledge necessary to construct a mapping function to accurately forecast the appropriate label for data sets known as testing data (unseen inputs). Multinomial Naïve Bayes, Support Vector Classification, Logistic Regression, Neural Networks, Random Forest, AdaBoost, Gradient Boosting, and Decision Tree are classifier models of supervised learning algorithms for multiclass classification.

In unsupervised learning algorithms, on the other hand, the model learns from unlabeled data without any predefined target labels. The objective is to discover patterns, structures, or relationships within the data by automatically developing classification labels by searching for similarities between data pieces to determine the category and create groups or clusters [13]. Clustering algorithms, such as K-means clustering, Latent Dirichlet Allocation (LDA), and Principal Component Analysis (PCA), are common algorithms of unsupervised learning. In this research, Multinomial Logistic Regression is selected as classifier algorithms to predict sentiment into positive, negative, and neutral classes in sentiment analysis.

Ramadhan et al. [14] applied Multinomial Logistic Regression in social media for Jakarta Governor Election with K-Fold cross-validation, achieving 74% accuracy with 90:10 training and testing data ratio. In his study, features were extracted and transformed into binary vectors using the TF-IDF method with the training dataset labeled manually. In addition, Purwandari et al. [15] compared Multinomial Logistic Regression and Multinomial Naïve Bayes for classifying weather in social media into five classes: cloudy, sunny, rainy, heavy rain, and light rain. The Multinomial Logistic Regression model has higher performance, with an accuracy rate of 83.3% and a precision rate of 90.3%. In comparison, the Multinomial Naïve Bayes model achieves an accuracy rate of 73.5% and a precision rate of 86.3%. Multinomial Logistic Regression has proven effective in classifying weather with good results.

C. Hybrid Approach in Sentiment Analysis

The hybrid approach is a methodology that integrates a lexicon-based approach and a machine-learning approach. Research investigations have recently focused on implementing a hybrid approach due to their ability to produce improved performance over either the lexicon-based model or the machine learning-based approach alone. Combining the advantages of lexicon-based and machine-learning methodologies is the primary goal of utilizing a hybrid method. The lexicon-based method is effective and can be used in a variety of contexts. This method does not require a significant amount of human interaction to label the training dataset and its ability to find the opinion words with the specific content orientation. On the other side, the machine learning approach effectively discovers subjectivity issues, noise resistance, and the ability to analyze numerous categories [16].

Most of the work in sentiment analysis performs supervised machine learning in binary classification with the best algorithms achieving an accuracy of less than 90%. A study by Pang et al. [17] demonstrated that the Support Vector Machine was 82.9% more accurate than the Naïve Bayes method in a movie review with binary classification. Vyas & Uma [18] observed that Support Vector Machine outperforms the Naïve Bayes and Decision Tree for binary sentiment classification in social media with an accuracy of 82.61%. Moreover, a study by Gupta et al. [19] discovered that the Random Forest algorithm outperforms Naïve Bayes, Support Vector Machine, and K-Nearest Neighbor (KNN) with an accuracy rate of 78% for binary customer sentiment classification using Amazon, Yelp, and IMDB dataset.

There is a minimum study for a hybrid approach in multiclass sentiment classification, especially for VADER, a lexicon-based approach with a supervised machine learning classifier. Chaithra [20] analyzed the metadata of media-sharing sites (YouTube) with popular videos using a hybrid approach. The lexicon-based approach VADER was applied, and a Naïve Bayes classifier as machine learning was trained with 70% of the data. The classifier achieved an accuracy of 79.78% and an F1 Score of 83.72% on 30% testing data. By applying this approach, the accuracy value indicates that the text comment can be classified into positive or negative feedback rather than like/dislike on the YouTube site.

To add on, Mahmood et al. [21] employed a hybrid approach to classifying public opinion on social media in positive and negative sentiment, combining a lexicon-based with machine learning approaches such as Naïve Bayes and Support Vector Machines. The Support Vector Machine performs superior to the Naïve Bayes classifier, achieving an accuracy rate of 80% before combining with the lexicon-based approach, while the lexicon-based method alone reached 85%. The accuracy performance was increased after combining the lexicon-based and machine-learning approaches with a 90% accuracy rate.

According to the study by Rajeswari [4], observed multiclass classification using SentiWordNet with Logistic Regression for movie datasets achieved an accuracy of 89% compared to Naïve Bayes and K-NN classifiers. In addition, Mujahid et al. [22] applied a hybrid approach to classify tweet e-learning implementation using VADER, TextBlob, and SentiWordNet combined with Logistic Regression, SVM, K-NN, and Random Forest. Their study showed that VADER and Random Forest outperformed with an accuracy of 88%. Not only that, Sham and Mohamed [23] used a hybrid approach to classify sentiment into tri-class (positive, negative, neutral) using climate change tweets. The study found that hybrid approaches, including Logistic Regression and TextBlob using TF–IDF, outperformed on a combined dataset with an F1-score result of 75.3%. Lemmatization techniques are not recommended during the data preprocessing phase of lexicon approaches, as they can decrease the performance obtained. This study also found that

TF–IDF as the feature extraction technique outperformed Bag-of-Words (BoW) when used in Logistic Regression.

III. Proposed Methodology

This section provides an overview of the proposed methodology employed with a hybrid approach for multiclass sentiment analysis. The sequential execution of the proposed methodology is illustrated in Fig. 1.

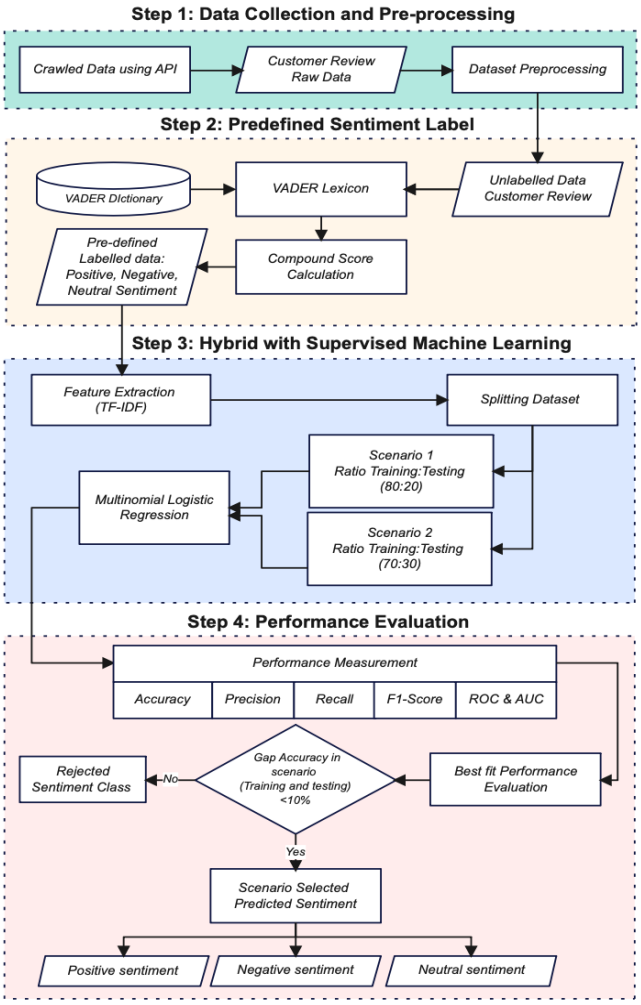


Fig. 1. Proposed methodology.

The first task is pre-processing, which includes transforming the case, deleting characters, and removing stop words to remove unneeded and repetitive information. Using the VADER lexicon, the second stage categorizes the customer review as positive, negative, or neutral. The third step is to use TF-IDF Vectorizer as a features extraction approach to execute Multinomial Logistic Regression for sentiment classification. The final step is to assess performance using accuracy, F1-Score, AUC, and ROC metrics.

A. Step 1: Dataset Collection and Pre-processing

1) Dataset: This study uses the Python programming language in a Jupyter Notebook to develop the suggested hybrid model. A web crawler was used to gather online customer reviews from Amazon.com, and mobile phone reviews were selected as a dataset with a more significant number of reviews. There are 3984 records of customer reviews with the following four attributes:

- Rating: consist of the customer assessment with range of 1 up to 5 stars to describe satisfaction level with the products. Rating on scale means: 1 - highly dissatisfied, 2 - dissatisfied, 3 - neutral, 4 - satisfied, and 5 - highly satisfied.
- Posting time: the date in customer post the review.
- Customer account: representing the customer identity in product review.
- Sentence of review: the text comments by customer after purchasing product to review the product performance.

All available information was crawled using a Python program. The Amazon website provides a review and rating score from 1-5 stars after customers buy products. Fig. 2 presents detailed rating information for the dataset.

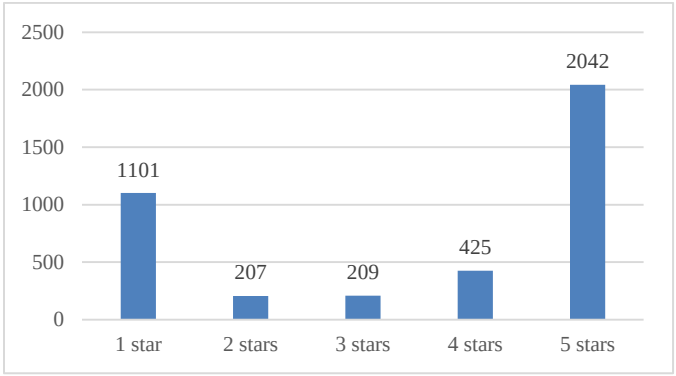


Fig. 2. Rating Information of mobile phone review

According to Fig. 2, the product has 1308 negative comments (33%) with 1-2 stars, 209 neutral sentiments (5%) with three stars, and 2467 with positive comments (62%) with 4-5 stars. This sentiment distribution indicates a positive product experience.

2) Pre-processing: Text pre-processing is important in sentiment analysis tasks due to the high dimensionality, poorly structured, and unstandardized text data tasks [24]. Pre-processing methods were performed to remove irrelevant and redundant data, which significantly impacted data quality. Pre-processing tasks include (1) case transformation, (2) tokenization, (3) stop word removal, and (4) stemming. Transform case is converted text to lowercase to ensure uniformity and prevent treating the same term differently due to case. The second procedure, tokenization, breaks textual data into smaller and meaningful components called tokens. Punctuation marks are removed during tokenization. The next step is to remove stop words. This method removes words from the text that don't add useful information, such as

determiners, prepositions, coordinating conjunctions, and more. Finally, stemming is applied to the reduction of words to their roots.

B. Step 2: Predefined Sentiment Label

The customer review dataset is labeled as positive, negative, or neutral using VADER vocabulary sentiment. This labeling procedure makes training data for the supervised machine-learning model more efficient. VADER lexicon provides sentiment strength based on the polarity scores for each data review with an intensity value that falls between the range of -1 to 1. For example, the words *"perfect"* and *"great"* will have the same positive polarity, whereas in the VADER lexicon, "great" is more positive than "perfect" with the intensity value (valence score) for *great (0.79)* higher than the intensity value of *perfect (0.69).*

The sentiment intensity analyzer in the NLTK package, known as VADER, produces a sentiment score in a dictionary with four terms: neg, neu, pos, and compound. The terms "neg," "neu," and "pos" are used to represent negative, neutral, and positive meanings accordingly. The total of the numbers should approximate or equal to 1. The compound sentiment score is determined by summing the valence scores of every word in the lexicon, and it serves as an indicator of the overall sentiment intensity. The emotion score ranges from (-1), indicating the most extreme negative sentiment, and (+1) the most extreme positive attitude. The experiment utilized the compound score threshold value to ascertain the inherent sentiment of a given text, as described in Table I and applied in Python using Algorithm 1.

TABLE I. VADER LEXICON COMPOUND SCORE

Sentiment Polarity	Range
Positive	Compound score ≥ 0.05
Negative	*Compound Score* $\leq$ -0.05
Neutral	-0.05 < compound score < 0.05

In order to implement the VADER algorithm in NLTK, it is necessary to download the VADER lexicon data and execute the commands using the Python script. The downloaded VADER lexicon was employed to apply the Sentiment Intensity Analyzer class from the NLTK for sentiment analysis. The Sentiment Intensity Analyzer class utilizes the polarity scores approach to provide a dictionary of sentiment scores. These scores include the compound score, a normalized composite value ranging from -1 to +1, and the positive, negative, and neutral scores. Based on specific requirements, the compound score threshold can be adjusted to categorize sentiment as positive, negative, or neutral.

Algorithm 1: The Sentiment VADER

Input: Pre-processed customer review data
Process:

```
import nltk
nltk.download('vader_lexicon')
from nltk.sentiment import SentimentIntensityAnalyzer
# Create an instance of the SentimentIntensityAnalyzer
analyzer = SentimentIntensityAnalyzer()
# Analyse sentiment of customer review sentence
sentence = "I love everything about the phone!"
sentiment_scores = analyzer.polarity_scores(sentence)
        if sentiment_scores['compound'] >= 0.05:
            sentiment = "positive"
        elseif sentiment_scores['compound'] <= -0.05:
            sentiment = "negative"
        else:
            sentiment = "neutral"
```

Output:

```
# Print sentiment scores
print sentiment_scores['compound']
```

C. Step 3: Hybrid with Supervised Machine Learning

The VADER-labeled data is paired with a supervised machine learning method employing Multinomial Logistic Regression to make predictions about the sentiment of customer reviews. Fig. 3 depicts the operational sequence of the Multinomial Logistic Regression model. The model is trained using a pre-defined VADER-labeled dataset. This dataset is divided into two different training and testing datasets scenario proportions.

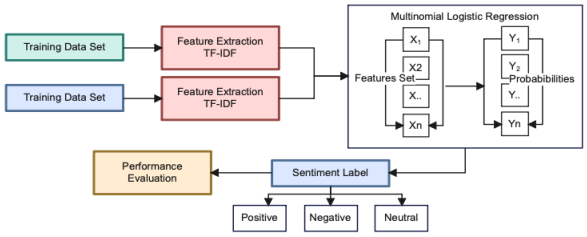


Fig. 3. Sentiment model with multinomial logistic regression.

1) Ratio proportion training and testing dataset: Training and testing data are split 80:20 and 70:30, respectively. In the first scenario, the 80% dataset is used for training and 20% for testing. In the second scenario, the 70% dataset is used for training and 30% is used for testing. This ratio was used when the dataset was large enough to provide sufficient training and testing instances. A larger training dataset can allow the model to learn more complex patterns, while smaller testing sets may result in less reliable performance estimates.

2) Feature extraction: This process aims to extract important and informative features from the cleaned dataset to reduce dimensionality and improve model performance and interpretability. The TF-IDF method was used in the feature extraction procedure to generate a document term matrix that denotes each term's word count and weightage [25].

$$\text{TF-IDF} = \text{tf x } 1 + \log\frac{N}{1+df(w)} \quad (1)$$

The term frequency **(tf)** is calculated as the value of that term's occurrence in specific documents. The total number of documents in the corpus is denoted by **(N)**, while **df(w)** signifies the count of documents containing the term **w**.

3) Sentiment classification model: There are two types of classification issues based on the number of classes: binary classification and multi-class classification (more than two classes). In this study, a neutral sentiment class was handled using multiclass classification. Multiclass classification enables finer-grained data analysis by considering many categories, resulting in a deep insight and understanding of the data.

By default, Logistic Regression is a classification algorithm for binary classification. The positive (true) class is allocated a value of 1, and the negative (false) class is assigned a value of 0. The fit model predicts the likelihood of a class 1 [26]. Multinomial Logistic Regression, or extended logistic regression, predicts more than two classes. Fig. 4 depicts the distinction between binary and multiclass classification.

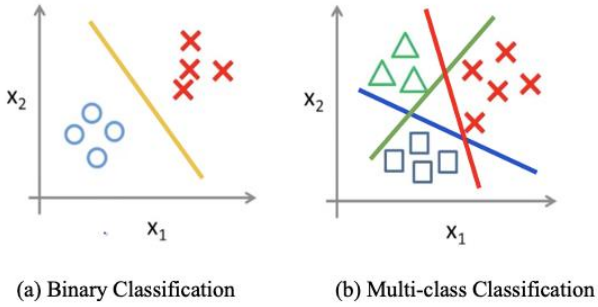


Fig. 4. Binary and multi-class classification.

A commonly used method for extending binary Logistic Regression to handle multiclass classification problems involves dividing the multiclass problem into several binary classification problems and applying a standard Logistic Regression model to each individual problem. This technique is called *one-vs-rest* and *one-vs-one* wrapper models. In the *one-vs-rest* or *one-vs-all* approach, build a Logistic Regression to find the probability the observation belongs to each class. Fig. 5 describes the step of multiclass Logistic Regression.

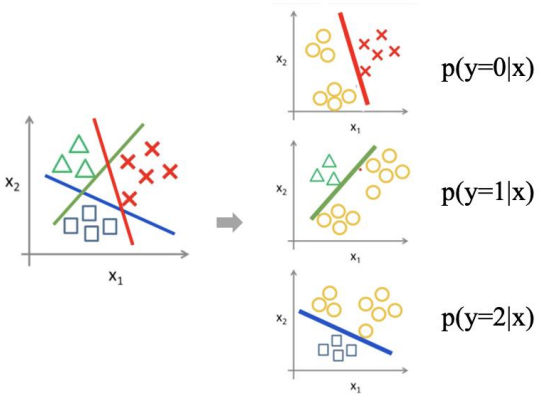


Fig. 5. Multiclass logistic regression classification.

The one-vs-rest technique involves training multiple binary Logistic Regression models, where each model represents one class against the rest of the classes [27]. The probabilities obtained from binary models combined to make multiclass classification predictions.

Denoted:

- N as the number of instances in the dataset
- K as the number of classes
- X as the input matrix of size N x D, where D is the number of features
- Y as the target variable matrix of size N x K, where each row represents the class labels for an instance using one-hot encoding

Train K binary logistic regression models, where each model i (i = 1 to K) predicts the probability of instance X belonging to class i against all other classes. The formula for the predicted probability of instance X belonging to class i is:

$$P(Y = i \mid X) = \sigma(W_i X + b_i) \quad (2)$$

where:

- $P(Y = i \mid X)$ is the predicted probability of instance X belonging to class i
- σ is the sigmoid function that maps the linear combination of the input features and model parameters to a probability between 0 and 1 (see Fig. 6).

$$\sigma(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+\exp(-z)} \quad (3)$$

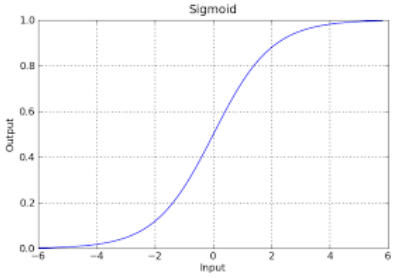


Fig. 6. Sigmoid function in logistic regression [27].

- W_i is the weight vector of size D x 1 for model i
- b_i is the bias term for model i

In order to generate predictions for a new instance X, the probabilities for each class must be computed utilizing the trained models, followed by the normalization probabilities with SoftMax formula [28]:

$$P(Y = i \mid X) = \frac{\exp(W_i X + b_i)}{\sum_{j=1}^{K} \exp(W_j X + b_j)} \quad 1 \le i \le K \qquad (4)$$

where:

- exp is the exponential function
- $\sum_{j=1}^{K} \exp(W_j X + b_j)$ is the sum of exponential terms for all classes j

Finally, the class label with the highest probability is assigned to the instance [29].

4) Model performance evaluation: The confusion matrix is a tabular representation that provides a detailed breakdown of a model's predictions. It displays the true positive, true negative, false positive, and false negative values for each class, allowing for an evaluation of the sentiment classification performance [30]. It helps to evaluate the correctness of classification approaches in multiclass classification problems. Table II presents the confusion matrix in multiclass classification with three sentiment categories: positive, negative, and neutral.

TABLE II. CONFUSION MATRIX FOR THREE SENTIMENT CLASSES

Actual	**Prediction**		
	Positive	***Negative***	***Neutral***
Positive	**TP**	**FNg1**	**FNt1**
	True Positive	False Negative 1	False Neutral 1
Negative	**FP1**	**TNg**	**FNt2**
	False Positive 1	True Negative	False Neutral 2
Neutral	**FP2**	**FNg2**	**TNt**
	False Positive 2	False Negative 2	True Neutral

where:

- True Positive (TP) refers to the number of times the classifier correctly predicts that the positive class is positive.
- The term True Negative (TN) refers to the number of the classifier that accurately predicts the negative class as negative.
- The term False Positive (FP) refers to the number of the classifier makes an incorrect prediction by classifying a negative class as positive.
- The number of the classifier incorrectly predicts the positive class as negative, referred to as False Negative(FN**).**

Performance evaluation metrics such as accuracy, precision, and recall can be formulated as follows by using a confusion matrix in Table II.

Accuracy is the percentage of correctly predicted instances across all classes divided by the total number of instances in the dataset.

$$Accuracy = \frac{TP+TNg+TNt}{TP+FP1+FP2+FNg1+TNg+FNg2+FNt1+FNt2+TNt} \; x\; 100\% \qquad (5)$$

Precision is the percentage of model positive predictions that are true. Precision in a multiclass classification problem can be calculated separately for each class, including positive, neutral, and negative.

$$Precision\ Positive = \frac{TP}{TP + FP1 + FP2} \; x\; 100\% \qquad (6)$$

$$Precision\ Negative = \frac{TNg}{TNg + FNg1 + FNg2} \; x\; 100\% \qquad (7)$$

$$Precision\ Neutral = \frac{TNt}{TNt + FNt1 + FNt2} \; x\; 100\% \qquad (8)$$

Recall called a true positive rate or sensitivity. Recall is also applied for each class in a multi-class classification problem that measures the percentage of true positive predictions out of all positive instances.

$$Recall\ Positive = \frac{TP}{TP + FNg1 + FNt1} \; x\; 100\% \qquad (9)$$

$$Recall\ Negative = \frac{TNg}{FP1 + TNg + FNt2} \; x\; 100\% \qquad (10)$$

$$Recall\ Neutral = \frac{TNt}{FP2 + FNg2 + TNt} \; x\; 100\% \qquad (11)$$

F1 Score is a metrics that combines precision and recall value. This value provides a balanced assessment of the model's performance for positive, negative, and neutral classes. The formula to calculate the F1 Score is as follows:

$$F1\ Score\ Positive = \frac{2\ x\ Precision\ Positive\ x\ Recall\ Positive}{Precision\ Positive + Recall\ Positive} \qquad (12)$$

$$F1\ Score\ Negative = \frac{2\ x\ Precision\ Negative\ x\ Recall\ Negative}{Precision\ Negative + Recall\ Negative} \qquad (13)$$

$$F1\ Score\ Neutral = \frac{2\ x\ Precision\ Neutral\ x\ Recall\ Neutral}{Precision\ Neutral + Recall\ Neutral} \qquad (14)$$

ROC Curve and AUC is a graphical representation of the model's performance across different classification thresholds.

IV. EXPERIMENT AND RESULT

This section presents the result of the experiments and evaluates the performance of VADER and Multinomial Logistic Regression.

A. VADER Result

The polarity results of the dataset are determined using the VADER vocabulary as presented in Table III, and an inconsistency rating was identified in the dataset review compared with the VADER result. The result of the VADER classification sentiment in positive, negative, or neutral was described in Fig. 7 compared with the distribution frequency of sentiment based on the rating score in the raw dataset. The VADER lexicon can detect the neutral sentiment by assigning the sentiment score a neutral range for a text that does not strongly express positive or negative sentiment.

TABLE III. VADER LEXICON COMPOUND SCORE

Sentence	Rating	Compound Score	VADER Polarity	Consistency
I love everything about the phone! It's in great condition	5	0.819	Positive	True
Bought for nephew	5	0.000	Neutral	False
Too difficult to set up	1	-0.3612	Negative	True
Like the style but soon as I turn it on it gets really hot	2	0.3612	Positive	False
It is a good phone but not a great one	3	0.000	Neutral	True
Battery capacity was at 85% when I got it. Disappointing	3	-0.4939	Negative	False

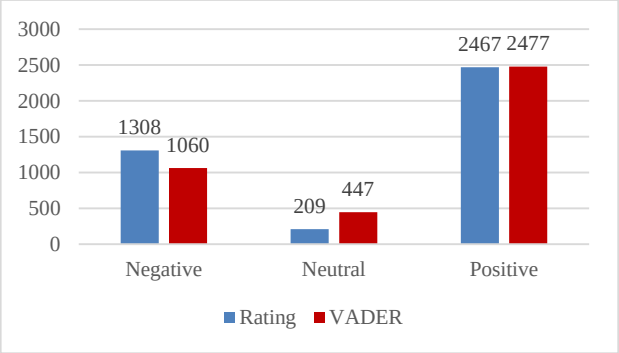


Fig. 7. Sentiment polarity distribution.

The accuracy rate for the VADER lexicon is 75.213% to label customer review data with sentiment polarity. VADER has been shown to perform effective in sentiment analysis tasks, particularly for customer reviews, where sentiments are often expressed in an informal and context-dependent manner.

TABLE IV. VADER LEXICON COMPOUND SCORE

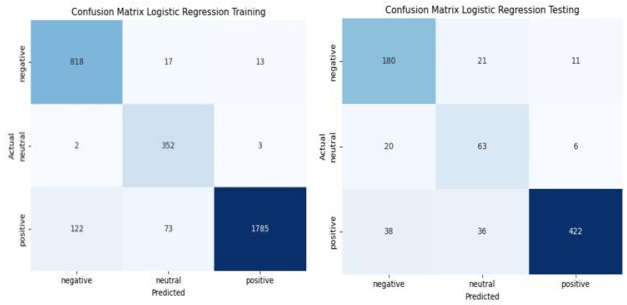
Bias Category from Rating to VADER sentiment	Total
Positive to Negative	**308**
Positive to Neutral	201
Negative to Positive	**263**
Negative to Neutral	206
Neutral to Positive	102
Neutral to Negative	**67**

According to Table IV, the investigation indicates misclassification between positive and negative and minimal misclassification between negative from neutral across all lexicons. VADER has a bias for inverting the polarity of ratings from positive to negative and vice versa. Since the differentiation between neutral with positive and neutral with negative are the significant factors to be clear in investigating the sentiment customer review, VADER lexicon is appropriate to adapted in task of labelling sentiment related to customer review.

B. *Sentiment Classification with Hybrid Approach: VADER with Multinomial Logistic Regression*

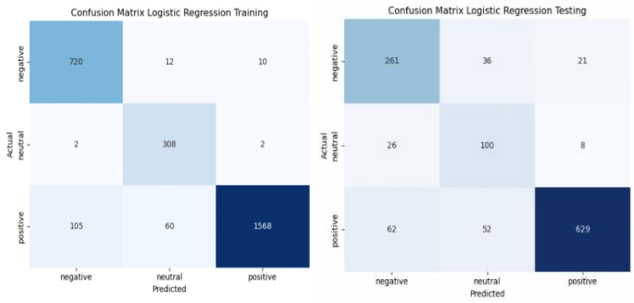
The metrics accuracy, precision, recall, F1-score, and Area Under Curve (AUC) are employed to assess the performance VADER with Multinomial Logistic Regression as a hybrid proposed model for multiclass classification. Confusion matrix visually shows the performance with two scenarios: splitting ratio 80:20 (see Fig. 8) and ratio 70:30 (see Fig. 9).

Regarding the result comparison in Table V, the first scenario with an 80:20 ratio is outperformed with an accuracy of 9,341% with a gap performance of <10% between training and testing data. Extending VADER with Multinomial Logistic Regression has increased the accuracy value by 17.565% from 75.213% with VADER to 92.778% with a hybrid approach. The second scenario has an overfitting condition with training data having a good performance, while its performance decreases significantly on the testing dataset (new data). Overfitting in machine learning should be avoided because it can negatively impact the model's performance and generalization ability on data that has never been seen before.



(a) Training data (b) Testing Data

Fig. 8. Confusion matrix with ratio 80:20.



(a) Training data (b) Testing Data

Fig. 9. Confusion matrix of with ratio 70:30.

TABLE V. GAP ANALYSIS FOR ACCURACY PERFORMANCE

Ratio	Dataset	True Predicted			Accuracy (%)	Gap Performance (%)
		Pos	*Neg*	*Neu*		
80:20	Training	1785	818	352	92.778	9,341
	Testing	422	63	180	83.437	
70:30	Training	1568	720	308	93.147	10.293
	Testing	629	261	100	82.845	

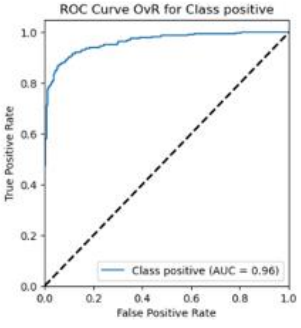
According to Table VI, the F1-Score for the positive, negative, and neutral classes are greater than 0.5 and close to 1. The positive classes have a higher score compared to the other classes. F1-Score combines precision and recall value works where the dataset is imbalanced, and according to this metric performance, a hybrid approach with a combination of VADER lexicon and Multinomial Logistic Regression accurately identifies positive, negative, and neutral classes with high recall and minimizes the frequency of false positives with high precision.

Furthermore, the AUC values in Table VI, which are close to 1, indicate that the hybrid approach is highly effective with excellent discrimination capabilities in distinguishing between positive, negative, and neutral classes. To better understand these data, the area under the ROC curve in Fig. 10 is represented by the area under the curve (AUC).

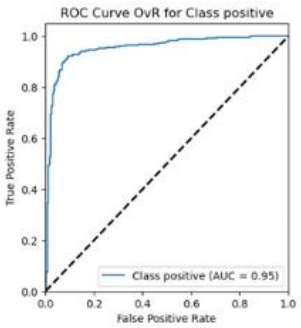
The ROC (Receiver Operating Characteristic) in Fig. 10 depicts a trade-off between the True Positive Rate and the False Positive Rate, which are shown on the "y-axis" and "x-axis," respectively. The line in the upper left corner of each ROC curve shows the cutoff value. The representation of ROC in Fig. 10 indicates that a hybrid approach with a combination of VADER lexicon and Multinomial Logistic Regression delivers a greater True Positive Rate (TPR) while preserving the low value in FPR.

TABLE VI. HYBRID APPROACH PERFORMANCE EVALUATION WITH SELECTED RATIO 80:20

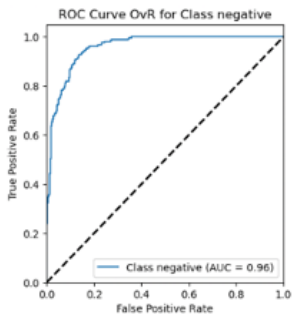
Dataset		Evaluation Metrics			
		Precision	***Recall***	***F1-Score***	***AUC***
Training	Positive	0.991	0.902	0.944	0.96
	Negative	0.868	0.964	0.914	0.96
	Neutral	0.796	0.984	0.881	0.90
	Macro Avg	0.885	0.951	0.913	
	Weighted Avg	0.934	0.928	0.929	
Testing	Positive	0.961	0.851	0.903	0.95
	Negative	0.756	0.849	0.800	0.95
	Neutral	0.525	0.708	0.603	0.91
	Macro Avg	0.748	0.803	0.769	
	Weighted Avg	0.858	0.834	0.842	



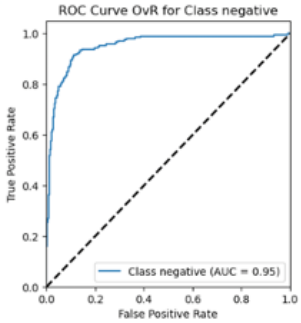
(a) Training Positive Class



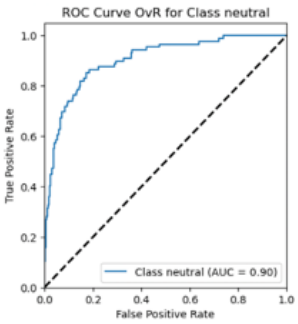
(b) Testing Positive Class



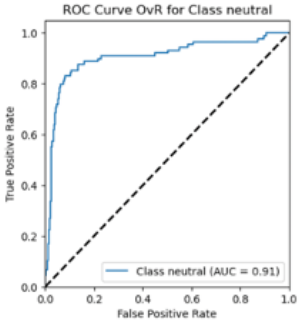
(c) Training Negative Class



(d) Testing Negative Class



(e) Training Neutral Class



(f) Testing Neutral Class

Fig. 10. Hybrid approach ROC visualization (Ratio 80:20).

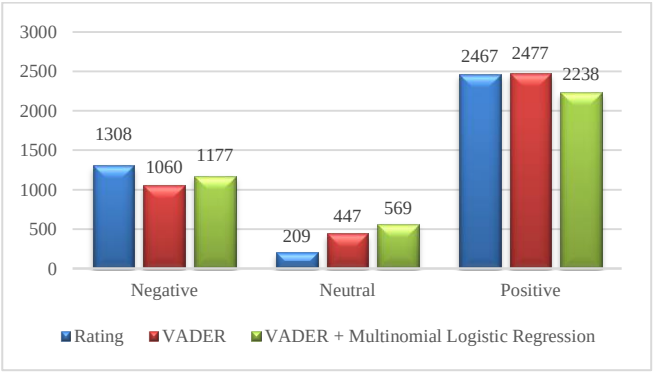


Fig. 11. Frequency distribution of sentiment polarity.

Fig. 11 presents the frequency distribution of sentiment polarity as determined by the rating score, VADER lexicon, and hybrid approach. The comparison results demonstrate that the hybrid approach effectively handles the neutral class when applied to multiclass sentiment classification. Extending VADER with Multinomial Logistic Regression can classify customer sentiment in online reviews as positive, negative, and neutral with good performance. The effectiveness of this method depends on the dataset accessibility, the complexity of sentiment nuances to be captured, and the specific criteria in the domain application.

V. CONCLUSION AND FUTURE WORKS

Online customer reviews can accurately predict customer sentiment regarding product experiences after purchase. With an accuracy percentage of 75.213%, VADER Lexicon classifies customer evaluations as positive, negative, or neutral, efficient in terms of time and costs for text labeling

without a human annotator. The VADER model is interpretable and simple to use, but its functionality is limited by the lexicon. This method depends on words in the dictionary and may only work well with words in the lexicon. Additionally, the accuracy increased to 92.778% after combining with Multinomial Logistic Regression. The high level of accuracy indicates that the hybrid approach has an excellent performance in predicting the sentiment polarity in multiclass classification in customer reviews. The VADER model performs well in predicting the neutral class, whereas the Multinomial Logistic Regression model succeeds in an imbalanced dataset with high-dimensional features and overfitting challenges. Further investigation could involve conducting experiments to fine-tune the hybrid approach and comparing it to various algorithms in lexicon-based and machine-learning approaches.s

REFERENCES

[1] K. Jindal and R. Aron, "A Systematic Study of Sentiment Analysis for Social Media Data," Materials Today Proceeding, Feb. 2021.

[2] H. AL-Rubaiee, R. Qiu, and D. Li, "The Importance of Neutral Class in Sentiment Analysis of Arabic Tweets," International Journal of Computer Science and Information Technology, vol. 8, no. 2, pp. 17–31, Apr. 2016.

[3] N. Hu, P. A. Pavlou, and J. Zhang, "On self-selection biases in online product reviews," MIS Quarterly, vol. 41, no. 2, pp. 449–471, 2017.

[4] A. M. Rajeswari, M. Mahalakshmi, R. Nithyashree, and G. Nalini, "Sentiment Analysis for Predicting Customer Reviews using a Hybrid Approach," in Proceedings - Advanced Computing and Communication Technologies for High Performance Applications, Institute of Electrical and Electronics Engineers Inc., Jul. 2020, pp. 200–205.

[5] T. Sana, B. Ines, J. Salma, and Y. Ben Ayed, "A Hybrid Method for Arabic aspect-Based Sentiment Analysis," International Journal Hybrid Intelligence System, vol. 16, no. 2, pp. 99–110, 2020.

[6] V. U. Ramya and K. T. Rao, "Sentiment Analysis of Movie Review using Machine Learning Techniques," International Journal of Engineering & Technology, vol. 7, no. 2, pp. 676–681, 2018.

[7] S. Singh Hanswal, A. Pareek, and A. Sharma, "Twitter Sentiment Analysis using Rapid Miner Tool," 2019.

[8] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," in Proceedings-International AAAI Conference on Weblogs and Social Media, Association for the Advancement of Artificial Intelligence, 2014, pp. 216–225.

[9] S. Panchal, (2020, March 7), "Sentiment Analysis with VADER- Label the Unlabelled Data," Medium Website, Analytics Vidhya. https://medium.com/analytics-vidhya/sentiment-analysis-with-vader-label-the-unlabeled-data-8dd785225166

[10] M. A. Al-Shabi, "Evaluating The Performance of The Most Important Lexicons Used to Sentiment Analysis and Opinions Mining," International Journal of Computer Science and Network Security, vol. 20, no. 1, pp. 51–57, 2020.

[11] D. Heaton, J. Clos, E. Nichele, and J. Fischer, "Critical reflections on three popular computational linguistic approaches to examine Twitter discourses," PeerJ Computer Science, vol. 9, p. e1211, Jan. 2023.

[12] V. Nasteski, "An Overview of The Supervised Machine Learning Methods," University St. Kliment Ohridski - Bitola, Dec. 2017.

[13] T. O. Ayodele, "Types of Machine Learning Algorithms," InTech, Feb. 2010.

[14] W. P. Ramadhan, A. Novianty, and C. Setianingsih, "Sentiment Analysis Using Multinomial Logistic Regression," in 2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC), IEEE, Sep. 2017, pp. 46–49.

[15] K. Purwandari, T. W. Cenggoro, J. W. C. Sigalingging, and B. Pardamean, "Twitter-based Classification for Integrated Source Data of Weather Observations," International Journal of Artificial Intelligence, vol. 12, no. 1, pp. 271–283, Mar. 2023.

[16] F. Hemmatian and M. K. Sohrabi, "A Survey on Classification Techniques for Opinion Mining and Sentiment Analysis," Artificial Intelligence Review, vol. 52, no. 3, pp. 1495–1545, Oct. 2019.

[17] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," in Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2002, pp. 79–86.

[18] V. Vyas and V. Uma, "An Extensive study of Sentiment Analysis tools and Binary Classification of tweets using Rapid Miner," in Procedia Computer Science, Elsevier B.V., 2018, pp. 329–335.

[19] K. Gupta, N. Jiwani, and N. Afreen, "A Combined Approach of Sentimental Analysis Using Machine Learning Techniques," Revue d'Intelligence Artificielle, vol. 37, no. 1, pp. 1–6, Feb. 2023.

[20] V. D. Chaithra, "Hybrid Approach: Naive Bayes and Sentiment VADER for Analyzing Sentiment of Mobile Unboxing Video Comments," International Journal of Electrical and Computer Engineering, vol. 9, no. 5, pp. 4452–4459, 2019.

[21] A. T. Mahmood, S. S. Kamaruddin, R. K. Naser, and M. M. Nadzir, "A Combination of Lexicon and Machine Learning Approaches for Sentiment Analysis on Facebook," Journal of System and Management Sciences, vol. 10, no. 3, pp. 140–150, 2020.

[22] M. Mujahid et al., "Sentiment Analysis and Topic Modelling on Tweets about Online Education During Covid-19," Applied Sciences, vol. 11, no. 8348, pp. 2–25, Sep. 2021.

[23] N. M. Sham and A. Mohamed, "Climate Change Sentiment Analysis Using Lexicon, Machine Learning and Hybrid Approaches," Sustainability, vol. 14, no. 4723, pp. 1–28, Apr. 2022.

[24] C. Zhu and D. Gao, "Influence of data pre-processing," Journal of Computing Science and Engineering, vol. 10, no. 2, 2016.

[25] M. R. Hasan, M. Maliha, and M. Arifuzzaman, "Sentiment Analysis with NLP on Twitter Data," in 5th International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering, Institute of Electrical and Electronics Engineers Inc., Jul. 2019.

[26] J. Brownlee, (2020, September 1), "Multinomial Logistic Regression With Python," Machine Learning Mastery. Accessed: Oct. 10, 2023. https://machinelearningmastery.com/multinomial-logistic-regression-with-python/

[27] "Introduction to Machine Learning Multi-class Classification," PowerPoint slides. 2020. Carnegie Melon University.

[28] J. Daniel and J. H. Martin, "Logistic Regression," in Speech and Language Processing, 2023, pp. 1–25.

[29] C. Wakamiya, "Classification with Logistic Regression," PowerPoint slides. 2020. Berkeley SCET.

[30] A. E. S. Saputro, K. A. Notodiputro, and Indahwati, "Study of Sentiment of Governor's Election Opinion in 2018," International Journal of Scientific Research in Science, Engineering , and Technology, vol. 4, no. 11, pp. 231–238, Dec. 2018.