JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 2, Juni 2025, Pp. 1257-1271

SENTIMENT CLASSIFICATION IN E-COMMERCE USING NAÏVE BAYES AND COMBINED LEXICON - N-GRAM FEATURES

Nabiel Muhammad Al Ghazali*[1], Yuliant Sibaroni[2]

1. Telkom University, Indonesia
2. Telkom University, Indonesia

Article Info

Keywords: Sentiment Analysis, E-commerce, Naïve Bayes, Lexicon-based Features, N-gram Features, Feature Extraction, Machine Learning, Natural Language Processing, Shopee, Customer Reviews

Article history:
Received 11 November 2024
Revised 15 Desember 2024
Accepted 14 Januari 2024
Available online 1 Maret 2025

DOI :
https://doi.org/10.29100/jipi.v10i2.6157

* Corresponding author.
Nabiel Muhammad Al-Ghazali
E-mail address:
ghaziel@student.telkomuniversity.ac.id

ABSTRACT

This study investigates sentiment classification in e-commerce using Naïve Bayes with lexicon-based, N-gram, and combined lexicon-N-gram features. While previous research has employed various e-commerce platforms and achieved varying degrees of accuracy using Naïve Bayes for sentiment analysis, the combination of lexicon and N-gram features with Naïve Bayes has not been extensively explored in e-commerce contexts. This study proposes to evaluate three models: Naïve Bayes with Lexicon Features, Naïve Bayes with N-Gram Features, and Naïve Bayes with Combined Lexicon-N-Gram Features. The research analyzes 10,000 customer reviews of the Shopee application from the Google Play Store. Results show that the Naïve Bayes model using combined lexicon-N-gram features achieved the highest performance among the three approaches. Using 10-fold cross-validation, the combined model achieved an average accuracy of 83.4%. The N-gram model showed strong performance with an average accuracy of 82.8%, while the lexicon-based model demonstrated lower performance with an average accuracy of 77%. These findings contribute to the field of sentiment analysis in e-commerce, highlighting the effectiveness of combining lexicon and N-gram features when used with Naïve Bayes classifiers. The study provides insights into optimizing sentiment classification techniques for e-commerce platforms, emphasizing the importance of leveraging both semantic and contextual information in sentiment analysis tasks.

I. INTRODUCTION

Sentiment analysis is a technique of computationally recognizing and classifying the viewpoints represented in a text, to determine the writer's opinion whether positive or negative [1]. It has been helpful in practically every business and social domain because the customer sentiment about the products and services of a business is determined by the comments and reviews [2]. Sentiment categories such as positive and negative, allow users to select opinions according to their preferences [3]. With so many posts on social media, sentiment analysis becomes crucial to classify structured and unstructured data [4]. Online shopping is now an unavoidable necessity for people, and electronic commerce (e-commerce) refers to the electronic exchange of goods and services involving electronic communication tools such as telephones and the internet [5]. One of the e-commerce in Indonesia that has the largest users is Shopee.

Several previous studies have conducted sentiment analysis using various research methods. In research [6] on sentiment analysis of e-commerce review that target to Shopee using Naïve Bayes Classifier with TF-IDF Weighting, the results show that the accuracy value is 72%, recall value is 72% and precision value is 78%. Then, in research [7] analyzed comparison of Naïve Bayes and Logistic Regression in sentiment analysis on marketplace reviews using Rating-Based Labeling, the results of the Naïve Bayes testing with 2 label datasets for Tokopedia is precision value of 81.13%, a recall value of 81.03%, and an accuracy value of 80.88%. In research [8], analyzed the product review sentiment analysis at online store Jiniso Official Shop using Naïve Bayes Classifier (NBC) method using WordCloud. The accuracy rate of the sentiment analysis using NBC is 94%, or 0.941. Then, in research [9] conducted a comparative study of Support Vector Machine and Naïve Bayes Classifier with TF-IDF for sentiment analysis on Amazon product reviews, the test results showed that the accuracy of the naïve bayes classifier is 84%, the precision is 82.8%, the recall is 82.8%, and the f1-score is 82.6%. In research [10], applying the implementation of Naïve Bayes algorithm with TF-IDF Weighting and Unigram Feature on sentiment analysis application on Zalora and Berrybenka online stores, an accuracy of 86.66% is achieved for training data.

Based on the research results that have been described, there have been studies that analyze sentiment using Naïve Bayes for various e-commerce platforms with varying degrees of accuracy. These studies used various features and weighting methods, such as TF-IDF [6] [9], Rating-Based Labeling [7], and WordCloud [8]. The use of TF-IDF as a weighting method seems to be quite common and effective which achieved 86.66% accuracy [10]. However, there is no research that specifically combines Lexicon and N-gram features in sentiment analysis using Naïve Bayes for the Shopee platform. Therefore, the author wishes to continue the previous research approach in the hope that it can provide significant benefits for future research.

Previous research has employed various e-commerce platforms with varying degrees of accuracy that analyze sentiment using Naïve Bayes. While these methods have shown promise, they often require large datasets or significant computational resources. Naïve Bayes offers a balance of efficiency and effectiveness, particularly for smaller datasets. However, the combination of lexicon and N-gram features with Naïve Bayes has not been extensively explored in e-commerce contexts. Therefore, this study proposes to evaluate three models: Naïve Bayes with Lexicon Features, Naïve Bayes with N-Gram Features, and Naïve Bayes with Combined Lexicon-N-Gram Features. This new approach aims to address limitations of previous studies by leveraging the complementary strengths of lexicon and N-gram features to improve sentiment classification accuracy, enhancing context sensitivity, balancing precision and recall, addressing challenges like negations and intensifiers, and optimizing Naïve Bayes performance for smaller, specific e-commerce datasets. By combining these features, the research aims to achieve significant accuracy improvements compared to single feature type methods, providing new insights into optimizing sentiment classification techniques for e-commerce platforms.

Naïve Bayes was chosen as the primary algorithm for this research due to several key advantages in the context of sentiment analysis for e-commerce reviews. Firstly, Naïve Bayes is known for its efficiency and speed, particularly with large datasets, making it well-suited for processing the high volume of reviews common in e-commerce platforms. Secondly, it performs well with high-dimensional data, which is characteristic of text data when using bag-of-words or N-gram representations. Thirdly, Naïve Bayes has shown robust performance in text classification tasks, often achieving competitive results despite its simplistic assumptions.

When compared to other popular algorithms in sentiment analysis, such as Support Vector Machines (SVM) or Deep Learning models like LSTM, Naïve Bayes offers distinct advantages for this research. Unlike SVM, which can be computationally intensive for large datasets, Naïve Bayes scales well with data size, making it more practical for real-time or resource-constrained applications. While Deep Learning models can potentially capture more complex patterns, they typically require larger datasets and more computational resources for training. In contrast, Naïve Bayes can perform effectively even with smaller datasets, which is beneficial when working with specific subsets of e-commerce reviews.

Moreover, the interpretability of Naïve Bayes is a significant advantage in this research context. Unlike 'black box' models, the probabilistic nature of Naïve Bayes allows for easy interpretation of feature importance, which is crucial for understanding the factors influencing sentiment in e-commerce reviews. This interpretability aligns well with the goal of not just predicting sentiment, but also gaining insights into the linguistic patterns and key terms that drive customer opinions in the e-commerce domain. Lastly, the simplicity of Naïve Bayes makes it an excellent baseline model for comparing the effectiveness of different feature extraction methods (lexicon-based, N-gram, and combined approaches) in sentiment analysis. Its performance can serve as a benchmark for evaluating the impact of these feature engineering techniques, providing clear insights into their relative effectiveness in capturing sentiment in e-commerce reviews.

The choice to combine Lexicon and N-gram features in this research stems from the complementary strengths of these two approaches in sentiment analysis. Lexicon-based methods rely on predefined sentiment dictionaries, providing a solid foundation for identifying sentiment-bearing words and capturing domain-specific terminology. However, they may struggle with context-dependent sentiments and new expressions. N-gram features, on the other hand, capture local context by considering word sequences, making them adept at identifying sentiment patterns in word combinations - particularly valuable in e-commerce reviews. By combining these feature sets, we aim to leverage their strengths while mitigating weaknesses. This combination is expected to improve accuracy by enhancing context sensitivity, improving coverage of both established and emerging expressions, balancing precision and recall, and addressing negations and intensifiers that might be missed by simple lexicon lookups.

Previous research has shown promising results with either lexicon-based or N-gram approaches, especially in e-commerce reviews. By integrating these features and applying them to the Naïve Bayes classifier, known for its efficiency in text classification, we anticipate achieving higher accuracy compared to single feature type methods. This research aims to quantify the accuracy improvement achievable through this feature combination, providing valuable insights for future sentiment analysis tasks in e-commerce and potentially other domains. The approach ties directly to the context of e-commerce reviews, making it highly relevant to our specific research focus on Shopee application reviews.

Sentiment classification for the Shopee e-commerce platform will be performed utilizing the Naïve Bayes algorithm. In this research, the use of combined Lexicon and N-gram features will be tested to obtain an effective feature set for improving the accuracy of Naïve Bayes. Lexicon and N-gram combined features are expected to provide comprehensive lexical information as well as word occurrence patterns so that Naïve Bayes performance is more optimal in sentiment classification. The goal is to classify people's sentiment on Shopee reviews with higher accuracy by utilizing the advantages of these combined features.

In addition, this research also opens opportunities for further exploration, such as the comparison of Naïve Bayes performance with other machine learning algorithms, the development of more sophisticated preprocessing techniques, the application of ensemble methods, multi-label sentiment analysis, and the handling of cultural and regional language contexts in e-commerce reviews. The research results are expected to provide new insights related to improving the accuracy of Naïve Bayes in text sentiment classification and pave the way for innovation in sentiment analysis in the e-commerce industry.

II. Methods

This chapter describes the implementation processes strategy into practice to build on previous research that have been discussed. The dataset is derived through data scraping from Indonesian reviews of the Shopee application on the Google Play Store. After deriving the dataset, preprocessing data is executed, then data labeling is executed that specifically lexicon with lexicon-based and n-gram with manual data labeling. Then the Lexicon and N-gram feature extraction and followed by the feature combination, then the Naïve Bayes model training and lastly is the evaluation with metric evaluation and confusion matrix.

TABLE I.
OVERVIEW OF RESEARCH METHODOLOGY STAGES

Research Stage	Description
1. Dataset	Gathering user reviews from Google Play Store through scraping, resulting in 10,000 reviews.
2. Data Preprocessing	Cleaning and preparing the text for analysis with steps such as case folding, normalization, stemming, stopword filtering, tokenization, and translate.
3. Data Labeling	Labeling the data using two methods: VADER labeling and star score labeling.
4. Feature Extraction	Utilizing two main methods: TF-IDF Vectorizer for n-grams and lexicon-based method for word features.
5. Naïve Bayes Model Training	Training the Multinomial Naive Bayes model with three scenarios: Naïve Bayes with Lexicon, Naïve Bayes with N-Gram, and Naïve Bayes with Combined Lexicon and N-Gram Features.
6. Evaluation	Evaluating the model using accuracy, precision, recall, and F1-score metrics, as well as 10-fold cross-validation to ensure reliability and consistency of results.

Table I summarizes the key stages of the research methodology. The detailed processes that include these steps can be seen in Figure 1.

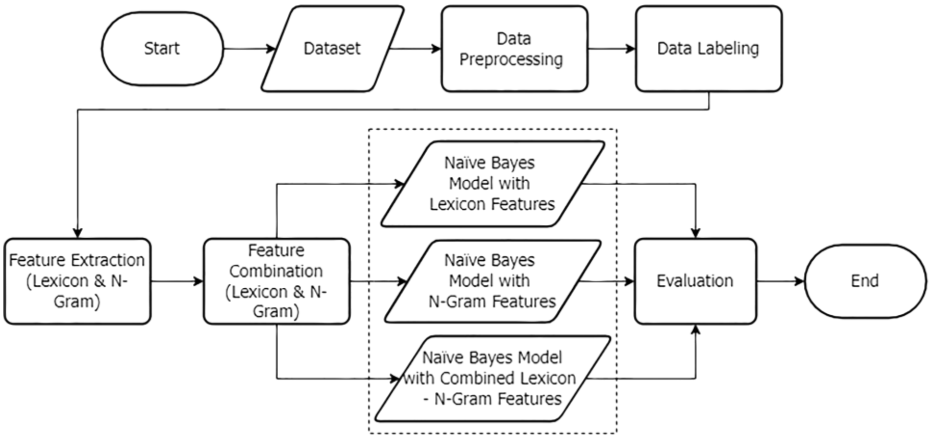
Start
Dataset
Data Preprocessing
Data Labeling
Feature Extraction (Lexicon & N-Gram)
Feature Combination (Lexicon & N-Gram)
Naïve Bayes Model with Lexicon Features
Naïve Bayes Model with N-Gram Features
Naïve Bayes Model with Combined Lexicon - N-Gram Features
Evaluation
End

Fig. 1. System Flow of the Processes

A. Dataset

The dataset utilized this study was derived through Shopee application reviews on Google Play Store. Dataset was collected as much as 10,000 review data. User reviews of the Shopee application serve as the source of the data. Data collection is done through a scrapping process and then stored in a csv file. Reviews that come from the Google Play Store provide an overview of the performance of an application. Some examples of datasets that are the result of data scraping can be seen in Table I.

TABLE II
RESEARCH DATASET

username	content	score
grils trilcee	Maaf saya rubah dulu bintangnya dari 5 ke 2. Karena semakin kesini semakin selalu diperbarui performa aplikasi bukan makin baik tapi makin buruk. Penonton saat live streaming drastis menurun, padahal jauh sebelum perbaruan" selalu stabil penontonnya. Jaringan juga selalu lag padahal stabil, coba live di tiktok dengan wifi & jaringan data yg sama selalu stabil. Tolong lah perbaiki lagi shopee yg katanya tempat belanja no. 1 ini. Supaya penjual dan pembeli lebih nyaman lagi bertransaksi.	2
Ani Rismayani	Selama menggunakan aplikasi ini ada banyak keuntungan yg saya peroleh, dapat berbelanja dengan harga yg cukup terjangkau ditambah ada banyak chasback dan discount serta gratis ongkir setiap bulan dan ada hadiah2 yg juga menarik	4
Ade Irmajayanti	Saya merasa sangat puas menggunakan Shopee. Dengan adanya keberagaman produk, penawaran eksklusif, dan diskon menarik, memberikan pengalaman berbelanja yang sangat memuaskan. Aplikasi ini mudah digunakan untuk membeli berbagai kebutuhan. Selain itu, dengan adanya program cashback memberikan kepuasan pada setiap penggunanya. Meskipun terdapat beberapa masalah dalam pengiriman, secara keseluruhan, saya merasa puas dan tetap memilih Shopee sebagai platform utama berbelanja online.	5
Ulil Amri Zach Ahmd	Minimal kalo mau nambahin fitur itu servernya dibenerin dulu, ngelag mulu apknya. Apalgi tu Shopee video, setiap mau ngeklik keranjangnya/komen kadang gk bisa padahal sinyal bagus. Upload video juga gitu, pas diupload bagus2 aja, eh pas diliat lagi jadi burik resolusi video ya. Trus mana pas abis update ngecrash lagi, vevek lah. Mau mencet bagian profil gk bisa, ngestuck kek hewan berkaki empat yang suka julurin lidah.	1
Fahmi Maulana	apk belanja fav gua, tapi ui kurang bagus, mana kadangâ€™ suka ngelag, gabisa dipencet, ngebug lah apalahh. semoga bisa ditingkatkan dan diperbaiki lagi sama developer nya. terima kasih	2

B. *Data Preprocessing*

Data preprocessing has an important role before carrying out in-depth analysis of text data. The purpose of preprocessing is to clean the data from noise and make it uniform in format for effective feature extraction. Several processes are involved in its data preprocessing, including translating dataset, Case Folding, Normalization, Stemming, Filtering Stopwords, and Tokenization which can be seen in Figure 2.

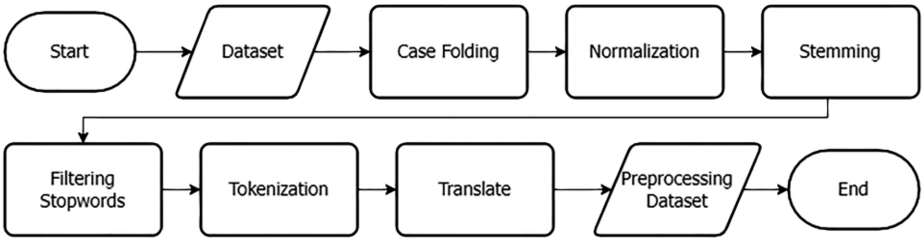

Fig. 2. Preprocessing Process

1. *Case Folding*

Case folding converts all characters to lowercase, ensuring uniformity across the dataset. This prevents variations in word casing from affecting the analysis.

2. *Normalization*

Normalization cleans the text by removing numerals, punctuation, special characters, and extra spaces. This process is implemented using Python's built-in string manipulation functions and regular expressions (re library). The punctuation list typically consists of characters such as commas (,), periods(.), exclamation marks(!), question marks (?), and other symbols that are not useful for sentiment analysis. This helps standardize the text format and makes it more suitable for analysis.

3. *Stemming*

Stemming reduces words to their root forms using the Sastrawi library, which is specifically designed for the Indonesian language. Sastrawi is an open-source library specifically designed for natural language processing tasks in the Indonesian language. For stemming, Sastrawi uses an effective algorithm to convert words to their base form, or root word. This process is important in text analysis because it can reduce the variety of words that come from the same form, such as nouns, verbs, or adjectives. By using Stemmer from Sastrawi, it can increase consistency in analysis and strengthen the results of the study, as the focus of the analysis can be placed on the more significant meanings of the root words.

4. *Stopwords Filtering*

Stopword filtering involves removing common and less meaningful words (stopwords) using the Sastrawi StopWordRemover. By using StopWordRemover from Sastrawi, can remove common words that lack meaning, such as "di," "yang," "atau," and "dan.". By filtering out these words, the analysis can focus on more significant terms that contribute to the overall meaning.

5. *Tokenization*

Tokenization involves splitting the text into individual words or tokens. This process is critical for analyzing the structure and meaning of the text. It allows for further processing, such as counting word frequencies or applying machine learning models.

6. *Translate*

The dataset which are mainly in Indonesian, were translated into English for this study aims to help sentiment analysis tools such as 'SentimentIntensityAnalyzer' that can provide more precise and reliable sentiment scores.

After going through this stage, the data is ready to be used for research process. This systematic preprocessing process is important to ensure the quality and relevance of the data in performing sentiment analysis.

C. Data Labeling

After going through data preprocessing, data labeling is the next process that goes through. Data labeling is used for the lexicon and n-gram feature extraction. The dataset in this study were labeled using two different ways based on VADER (Valence Aware Dictionary and sEntiment Reasoner) labeling that used for Naïve Bayes Model with Lexicon Features and the star rating score labeling that used for Naïve Bayes Model with N-Gram Features. For VADER labeling using a predefined dictionary then calculates the compound score that indicates the sentiment level of the sentiment level of the text and for the star rating score labeling initializing the sentiment manually based on the star scores given by users in their reviews. For VADER using a predefined sentiment dictionary, where words and phrases are given a sentiment score and using the VADER tool from the nltk library.

NLTK (Natural Language ToolKit) is a free open-source Python package that provides several tools for building programs and classifying data [11]. Equipped with a well-established sentiment lexicon, VADER is a rule-based tool for sentiment analysis [12]. VADER can determine sentiment scores by considering various factors, such as the text's vocabulary intensity and context. Then, the sentiment of each text is determined based on the compound value generated from the analysis. This compound score indicates the sentiment level of the text, where a positive value indicates a positive sentiment, and a negative value indicates a negative sentiment. In this research, the rule of VADER labeling is that if the compound value is greater than or equal to 0, the text is labeled "POSITIVE", and if the compound value is less than or equal to 0, the text is labeled "NEGATIVE" then the value of the text is negative and stored in a column of the dataset named 'sentiment_label'. The VADER labeling is used in the Naïve Bayes model with Lexicon Features.

TABLE III
LEXICON DATA LABELING

Compound	sentiment_label
0.8417	POSITIVE
0.9169	POSITIVE
-0.7684	POSITIVE
-0.3158	NEGATIVE
-0.5413	NEGATIVE

Meanwhile, the star rating score labeling has different ways with VADER, this labeling is executed manually based on the star scores given by users in their reviews. The star rating score labeling process starts by initializing the sentiment manually with the assumption that the score can reflect the sentiment of the text. Each text in the dataset is analyzed and labeled based on a subjective sentiment interpretation of the text content. The rule of star rating score labeling is that if the star rating score is 4 and 5, the text is labeled “POSITIVE”, and if the star rating score is 1 and 2, the text is labeled “NEGATIVE”. This manual data labeling process involves reading and evaluating each text to determine whether it is positive or negative and stored in a column of the dataset named 'label'. This star score labeling is used in the Naïve Bayes model with the N-Gram feature.

TABLE IV
N-GRAM DATA LABELING

Label	Count
POSITIVE	4504
NEGATIVE	3709

Table III shows the distribution of sentiment labels in the dataset used for the N-gram approach in sentiment analysis. This dataset consists of a total of 8,213 customer reviews that have been manually labeled based on their sentiments.

D. Feature Extraction (Lexicon and N-gram)

After data labeling is feature extraction from the dataset for lexicon and n-gram. Feature extraction is an important process in Natural Language Processing (NLP) that aims to convert raw text data into numerical representations that can be processed by machine learning models. The feature extraction process in this research aims to convert raw text into a numerical representation that can be understood by sentiment classification models, specifically using the Naïve Bayes approach. The feature extraction steps include using the 'TfidfVectorizer' method to measures the frequency of words by considering their importance in the document (TF-IDF).

Lexicon-based is a common method for performing sentiment analysis on social media due to its practicality. In a lexicon-based approach, the method uses a dictionary as a language or lexical source [13]. Lexicon Based is the characterization of words that have positive or negative sentiments based on dictionaries or lexicons. In the data labeling process, the Lexicon Based dictionary is used to calculate the sentiment score. After words with positive and negative sentiments are identified in a sentence, the next step is to calculate each word containing sentiment in the sentence by summing up the opinion values [14].

Lexicon feature extraction process starts by selecting words from a predefined sentiment dictionary, splitting the tokenized text data and lexicon sentiment labels from the dataset. Then, the data is split into train data to train the model and test data to measure model performance with a ratio of 80% and 20% using the 'train_test_split' function. Lexicon feature extraction is executed using 'TFidfVectorizer' from the scikit-learn library to convert text into TF-IDF vectors, but only for words that are in the sentiment dictionary. The lexicon feature extraction using 'TfidfVectorizer' by limiting with 'max_features' to 3,000 words due to the high accuracy. 'TfidfVectorizer' is used to convert text data into numerical vectors based on the term frequency-inverse document frequency (TF-IDF) method. This process produces a numerical representation of each document based on the occurrence of words in the text that has gone through the preprocessing stage.

This process yields lexicon features along with their TF-IDF values, which include positive and negative words relevant to sentiment. For example, positive words such as 'excellent' have a TF-IDF value of 0.82, 'satisfied' with 0.75, and 'happy' with 0.68. Meanwhile, negative words such as 'disappointed' have a TF-IDF value of 0.79, 'bad' with 0.71, and 'slow' with 0.65. Overall, this process generates 3,000 unique features specifically related to sentiment.

N-Grams are features used at the text processing stage in performing classification and sentiment analysis. N-Grams are used to combine frequently occurring words or give weight to important words in a document to indicate a sentiment [15]. The types of N-Grams applied in this study involve Unigram, which is a token consisting of one word; Bigram, which is a token consisting of two words; and Trigram, which is a token consisting of three words [15]. The n-grams used are not decomposed into small characters, but rather n-grams are used to break sentences into words [16].

N-gram feature extraction starts by splitting the tokenized text data and n-gram manual label from the dataset. Then, the data is split into train data to train the model and test data to measure model performance with a ratio of 80% and 20% using the 'train_test_split' function. N-gram feature extraction is executed using 'TfidfVectorizer' from the scikit-learn library to convert text into TF-IDF vectors based on on all words and phrases including unigrams, bigrams, and trigrams, so the text is converted into TF-IDF feature matrix for n-grams with the configuration 'ngram_range=(1, 3)' which means to generate sequential combinations of words from unigrams, bigrams, and trigrams in one feature and 'max_features' to 20,000 words due to the high accuracy. This approach allows the model to consider the order of words in the text, which can provide additional context in sentiment analysis.

The N-gram feature extraction process generates features along with their TF-IDF values, encompassing both common words and phrases that may or may not be directly related to sentiment. For example, the unigrams include words like 'product' with a TF-IDF value of 0.65, 'delivery' with 0.72, and 'price' with 0.80. The bigrams feature phrases such as 'very_satisfied' with a TF-IDF value of 0.92 and 'bad_quality' with 0.78. Additionally, trigrams include phrases like 'very_fast_delivery' with a TF-IDF value of 0.95. This process results in the generation of 20,000 unique features that encompass a wide range of common words and phrases, which may or may not be directly tied to sentiment.

The results of the lexicon and n-gram feature extraction for the train data and test data were evaluated by printing the extracted features in the form of arrays. Feature extraction is important in converting raw text into a numerical form that machine learning algorithms can understand, therefore helping to improve model efficiency and accuracy.

E. Feature Combination (Lexicon and N-Gram)

Lexicon and n-gram feature extraction are converted into Compressed Sparse Row (CSR) form using 'csr_matrix' function from 'scipy.sparse'. The CSR form was chosen due to its efficiency in storing and processing the sparse matrix, which is commonly generated from the text feature extraction. The CSR matrix for lexicon and n-gram features are subsequently combined using the 'hstack' function of 'spicy.sparse', which joins the two matrixes horizontally. This process results in one combined matrix containing information from both feature extraction methods. The combination of these two types of features in the Naïve Bayes model with the combined Lexicon-N-Gram feature results in a total of 23,000 features, which is expected to improve the accuracy of sentiment classification by utilizing the advantages of both approaches. This combined matrix is subsequently converted back to an array to facilitate the subsequent processes in the machine learning pipeline, namely model training. By combining features from lexicon and n-gram feature extraction, aims to improve accuracy and robustness in sentiment classification. The results show the combined matrix size for train data and test data, giving an indication of the scale and complexity of the data being processed.

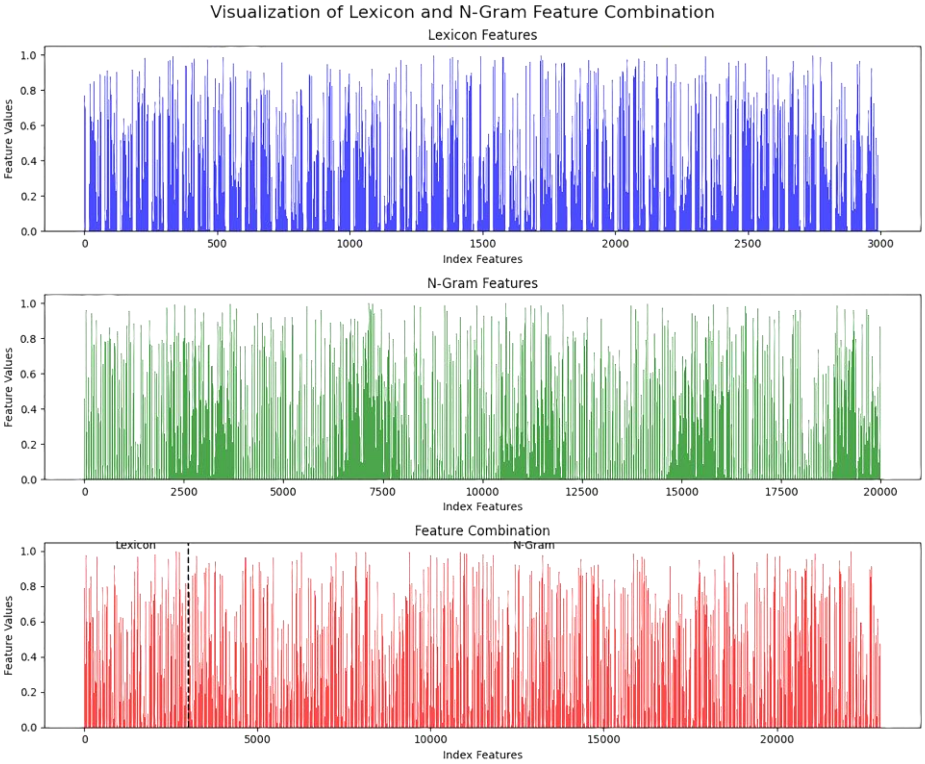

Fig. 3. Visualization of Feature Combination

Figure 3 illustrates the process of combining Lexicon and N-Gram features in sentiment analysis. The first graph displays the distribution of 3,000 Lexicon features, with values varying between 0 and 1, showing a relatively even pattern with few significant peaks. The second graph visualizes 20,000 N-Gram features, also with values ranging from 0 to 1, but displays a denser and more complex distribution, reflecting the diversity of linguistic patterns captured. The third graph illustrates the result of combining the two types of features, with a dotted line separating Lexicon features on the left and N-Gram features on the right.

F. Naïve Bayes Model Training

This section describes the naïve bayes classification model using two types of feature extraction, namely lexicon and n-gram, as well as a combination of both, to evaluate the performance of the model in performing sentiment analysis on text data. Naïve Bayes is a commonly used algorithm in sentiment analysis, invented by Tomas Bayes. The advantage of Naïve Bayes lies in the simplicity of the algorithm that is still able to provide a high level of accuracy [17]. Naïve Bayes is a type of model algorithm that is simple and easy to use and can effectively make predictions in a case based on text classification results [15].

Naïve Bayes will predict future opportunities based on previous experience, which is known as Bayes' Theorem. The main characteristic of Naïve Bayes Classification is a very strong prediction result (Naïve) of the independence of each condition or event. Conditional probability is the calculation of the probability of an event, A, when another event, B, has occurred, recorded as P(A|B), which combines the probabilities of A and B. This theory is used to measure the likelihood of a data set joining a particular category based on the available inferential data [18].

The advantage of using Naïve Bayes is that this algorithm does not require a lot of training data to select the parameters required in the classification process [18].

The Naïve Bayes process involves several steps, including calculation of the number of classes or labels, calculation of the number of probabilities per class, multiplication of all class variables, and comparison of products per class [18]. Bayes' theorem aims to calculate the probability of an event occurring based on other events that have occurred [18]. In simple terms, we assume that the presence of a word in a sentence is not influenced by other words. In sentiment analysis, each occurrence of a word has a weight that is calculated as the total weight for the entire sentence. This determines whether the sentence is positive or negative [19].

In this study, the Multinomial Naïve Bayes model is used, which is well suited for text classification tasks. The Multinomial model assumes that the features (words in the case) are generated from a multinomial distribution, which works well with the discrete nature of text data. The formula for Multinomial Naïve Bayes is in Equation (1) [20].

$$C_{best} = \arg\max_{y\,\in Y} \hat{P}(y) \prod_{j=1}^{m} \hat{P}(w_j|y) \tag{1}$$

Where C_{best} represents the predicted class label, either "positive" or "negative" sentiment. The arg max indicates the value of y that maximizes the expression inside the brackets. In this case, we are looking for the class (positive or negative) that gives the highest value. $y\ \in Y$ shows that Y is the set of all possible classes (positive and negative sentiments), and we are seeking the value of y within this set. $\hat{P}(y)$ is the prior probability of class y, the probability of a review being positive or negative based on the training data. In other words, it is the probability that a review is positive or negative regardless of the words in the review. The symbol $\prod_{j=1}^{m} \hat{P}(w_j|y)$ represents the product of conditional probabilities for each word (w_j) in the review, given class y. Here, m is the number of words in the review. Conditional probability $\hat{P}(w_j|y)$ indicates how likely a word is to appear in reviews belonging to class y. In essence, this formula calculates the product of the prior probability of a class and the probabilities of each word in the review for that class. This calculation is performed for both classes (positive and negative), and the class with the highest value is chosen as the prediction.

In this model training process, the Multinomial Naïve Bayes model is trained using the lexicon, n-gram, and the feature combination. The Multinomial Naïve Bayes model is used to learn the probability distribution of these features against positive and negative sentiment labels. Model evaluation is performed using standard metrics such as accuracy, precision, recall, and F1-score. Cross Validation using 10-fold was applied to ensure the reliability and consistency of the model's performance across dataset and the best accuracy of the model scenario is sought. The model evaluation results provide deep insight into the model's ability to predict positive and negative sentiments from text. Three scenarios are included in the research methodology utilized in this study including:

1) *Naïve Bayes Model with Lexicon Features*

In training the Naïve Bayes Model with Lexicon Features using lexicon feature extraction as input that the data labeling process uses VADER. Lexicon features are extracted using 'TfIdfVectorizer' that convert text into a numerical representation that can be processed by machine learning models by limiting with 'max_features' to 3,000 words due to the high accuracy followed by training the model using the training data.

2) *Naïve Bayes Model with N-Gram Features*

In training the Naïve Bayes Model with N-Gram Features using n-gram feature extraction is used as input that the data labeling process uses the star rating scores given by users in their reviews. N-gram features are extracted using 'TfIdfVectorizer' with range of n-grams (1, 3) to capture patterns of occurrence of words in the form of unigrams, bigrams, and trigrams and limiting with 'max_features' to 20,000 words due to the high accuracy followed by training the model using the training data.

3) *Naïve Bayes Model with Combined Lexicon - N-Gram Features*

In training the Naïve Bayes Model with Combined Lexicon - N-Gram Features using lexicon and n-gram feature extraction. Then, these two feature sets are then combined using 'hstack' from 'scipy.sparse' to create a richer representation of the text. This combined feature matrix is converted into an array and used as input for the Naïve Bayes model. This process aims to combine the strengths of lexicon-based analysis and n-gram patterns, thus improving the model's ability to recognize and classify text sentiment followed by training the model using the training data.

TABLE V
ILLUSTRATION OF FEATURE COMBINATION RESULT

Review ID	Lexicon Features 1	Lexicon Features 2	...	Lexicon Features 3000	N-Gram Features 1	N-Gram Features 2	...	N-Gram Features 20000
1	0.0	0.023	...	0.0	0.001	0.0	...	0.045
2	0.015	0.0	...	0.007	0.0	0.0	...	0.012
...	...	...	...	...	...	...	...	...
8213	0.0	0.01	...	0.0	0.0	0.03	...	0.0

Table V illustrates the structure of the combined feature matrix resulting from merging lexicon-based and N-gram features. The table shows how each review (represented by Review IDs 1 to 8213) is characterized by a combination of 3,000 lexicon features and 20,000 N-gram features, totaling 23,000 features. The prevalence of 0.0 values indicates the sparse nature of the matrix, which is typical in text analysis where not all features are relevant to every review. This comprehensive representation, combining both lexicon-based sentiment information and N-gram contextual patterns, serves as input for the Naïve Bayes model in the combined approach, aiming to enhance sentiment classification accuracy.

G. *Evaluation*

Evaluation is one of the processes in the system flow that aims to assess and determine the best algorithm by evaluating the performance of the Naïve Bayes sentiment classification model which requires testing the predicted data. To evaluate the performance of the Naïve Bayes sentiment classification model, we utilized several metrics derived from the confusion matrix, including accuracy and F1-score. Additionally, to ensure the reliability and consistency of our model's performance, we employed 10-fold cross-validation. In this process, the dataset was divided into 10 equal parts. The model was then trained on 9 parts and tested on the remaining part, with this process repeated 10 times so that each part had a chance to be the test set. Performance metrics were then calculated by averaging the results from all 10 iterations.

Begins by initializing a list to store the confusion matrix for each fold in the cross-validation process. For each of the ten folds, a mask is created to determine which data will be used in that particular fold. The Multinomial Naïve Bayes classification model is employed to make predictions using the ‘cross_val_predict’ function, which generates predictions based on the cross-validation framework. Following the predictions, the confusion matrix is calculated for each fold, reflecting the comparison between actual values and predicted values from the model, and the results are stored for further analysis. All the confusion matrices from each fold are then combined to yield a single aggregated confusion matrix, providing an overview of the model's performance across the entire dataset.

Important evaluation metrics such as accuracy, precision, recall, and F1-score are computed for each fold based on the values derived from the confusion matrix. After calculating the metrics for each fold, evaluation metrics are also computed for the combined confusion matrix, offering a comprehensive view of the model's overall performance in sentiment classification. By utilizing cross-validation and confusion matrices in the evaluation, this research ensures that the developed model is reliable and exhibits strong performance in identifying sentiments within text data. This approach helps mitigate the impact of data variability and provides a more robust estimate of the model's performance on unseen data. In this study, 10-fold cross-validation was applied to all three Naïve Bayes model scenarios (with Lexicon features, N-gram features, and their combination).

Confusion matrix is a table that shows the amount of actual testing data and the model's prediction results. Confusion matrix consists of 4 terms, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). True positive means positive data that is correctly predicted by the model, while true negative is negative data that is correctly predicted. Meanwhile, false positive refers to negative data that is incorrectly predicted as positive, and false negative is positive data that is incorrectly predicted as negative. Confusion matrix has a structure which can be found in Table V.

TABLE VI.
CONFUSION MATRIX

Predicted Value	**Actual Values**	
	Positive (1)	**Negative (0)**
Positive (1)	TP	FP
Negative (0)	FN	TN

To evaluate the performance of the Naïve Bayes sentiment classification model, several metrics will be used including accuracy, precision, recall, and F1-score. The following are the formulas for the metrics that will be used.

1) *Accuracy*

Accuracy measures the ratio of the number of correct predictions to the overall data. The formula for calculating Accuracy is found in formula (2).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

2) *Precision*

Precision measures the ratio of correct positive predictions to all positive detection results. The formula for calculating Precision is in formula (3).

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

3) *Recall*

Recall measures the ratio of correctly predicted positive data. The formula for calculating Recall is found in formula (4).

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

4) *F1-Score*

F1-Score is the harmonic mean of precision and recall. The formula for calculating F1-Score is found in formula (5).

$$F1\ Score = 2\left(\frac{Precision\ \times Recall}{Precision + Recall}\right) \tag{5}$$

In this study, the data analysis method involves a machine learning approach to identify patterns and trends in e-commerce data. This analysis comprises several key stages, from data preparation to model evaluation, with an emphasis on using appropriate machine learning techniques to optimize sentiment classification results. The data used in this study comes from a database of e-commerce customer reviews. The data processing involves cleaning the data, removing stopwords, and normalizing the text to ensure consistency and quality. The use of Python software with libraries such as NLTK and Pandas allows for efficient and large-scale data processing.

Feature extraction is carried out using two main approaches: ‘TfIdfVectorizer’ for n-grams and a lexicon-based method for word features. The ‘TfIdfVectorizer’ is used to convert text into numerical representations that reflect the importance of words in the context of n-grams, while the lexicon-based method relies on a dictionary of positive and negative words to measure word polarity. The classification model used in this study is Multinomial Naive Bayes, chosen for its advantages in handling text data with discrete features and the independent nature of features. This model is trained with three scenarios: Naive Bayes with Lexicon, Naive Bayes with N-Gram, and Naive Bayes with Combined Lexicon and N-Gram Features.

For model evaluation, accuracy, precision, recall, and F1-score metrics are used. The 10-fold cross-validation process is employed to avoid overfitting and ensure reliable and consistent evaluation results. The scikit-learn software is used for implementing cross-validation and calculating evaluation metrics due to its ease of use and comprehensive documentation. The selection of this analysis method is based on the need to accurately identify sentiment in customer reviews, considering the balance between false positives and false negatives. Python and scikit-learn are chosen for their reliability in data analysis and extensive support for various machine learning algorithms.

III. Result and Discussion

A. *Result*

In this chapter, testing has been conducted for feature extraction, feature combination, and Naïve Bayes model training for Shopee e-commerce sentiment classification on Google Play Store with lexicon and n-gram approaches. The class frequencies for the lexicon data labeling and n-gram data labeling have balanced data. Then, feature extraction and feature combination for model training are performed.

The three model scenarios of the Naïve Bayes model were executed aims to compare and evaluate the effectiveness of different feature extraction methods in sentiment analysis and to provide a comprehensive understanding of how different feature extraction methods affect the performance of Naïve Bayes in sentiment classification, ultimately contributing to the field of sentiment analysis in e-commerce. Then, 10-fold cross validation was executed to assess the robustness and generalization capability of each Naïve Bayes model scenario (Lexicon, N-gram, and Combined features), ensuring consistent performance across different subsets of the data. This process also provided a fair and comprehensive comparison between the three scenarios of the Naïve Bayes model. Each model's performance is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. This chapter aims to understand the effectiveness of each method, compare their performance, and draw insights from the results. The model evaluation of three scenarios of Naïve Bayes models can be seen in Table IV:

TABLE VII
NAÏVE BAYES MODEL TRAINING RESULT

Method	Accuracy	Precision	Recall	F1-Score
Naïve Bayes Model with Lexicon Features	77.9%	77.3%	77.3%	77.3%
Naïve Bayes Model with N-Gram Features	84.4%	84.5%	84.8%	84.3%
Naïve Bayes Model with Combined Lexicon – N-Gram Features	84.9%	85.2%	85.4%	84.9%

B. Discussion

The combined Lexicon and N-Gram model achieves the highest performance among the three methods. Its accuracy stands at 84.9%, with a precision of 85.2%, recall of 85.4%, and an F1-score of 84.9%. This suggests that combining features can capture a wider range of sentiment indicators, leading to improved performance. Both precision and recall are high, indicating effective sentiment identification. The F1-score matches the accuracy, showing robust performance. This indicates that the combined model effectively identifies patterns and sequences of words that influence sentiment, which might be missed by using only lexicon-based or N-Gram features. The best accuracy during cross-validation using 10-fold was 85%.

The Naïve Bayes model using N-Gram features also performs well. Its accuracy stands at 84.4%, with a precision of 84.5%, recall of 84.8%, and an F1-score of 84.3%. The use of N-Gram features (including unigrams, bigrams, and trigrams) captures more context from the text, leading to improved performance. Both precision and recall are high, indicating effective sentiment identification. The F1-score closely matches the accuracy, showing robust performance. The best accuracy during cross-validation using 10-fold was 84.6%.

The Lexicon model, while still performing well, shows slightly lower metrics. It achieves an accuracy of 77.9%, precision of 77.3%, recall of 77.3%, and an F1-score of 77.3%. Precision and recall values are close to the accuracy, suggesting balanced performance in identifying both positive and negative sentiments. This suggests that lexicon features, based on predefined sentiment dictionaries, might not capture the contextual nuances as effectively as N-Grams. The best accuracy during cross-validation using 10-fold was 80.5%.

The final evaluation metrics for the Naïve Bayes Model with Combined Lexicon-N-Gram Features are as follows: The accuracy of 83.4% reflects the model's overall performance in correctly classifying the sentiments of customer reviews. The precision of 89.5% indicates a high level of correctness in positive sentiment predictions, meaning most predicted positives are true positives. The recall of 79.1% shows that the model successfully identifies a substantial portion of actual positive sentiments. The F1-score of 84% balances the precision and recall, providing a comprehensive measure of the model's accuracy. These evaluation metrics highlight the effectiveness of combining lexicon and N-Gram features, demonstrating significant improvements in accuracy, precision, and recall, and F1-score. This balance is vital for practical applications in sentiment analysis, where accurate identification of positive sentiments is crucial, alongside minimizing false positives. The combination of these features enhances the model's ability to capture a wider range of sentiment indicators, ultimately leading to improved performance in sentiment classification tasks.

The performance improvement from the Lexicon model to the N-Gram model, and further to the combined Lexicon-N-Gram model, shows that the combination of features can capture more aspects of sentiment in the text. The combined model is able to leverage the strengths of both approaches, resulting in marginal but consistent improvements in all evaluation metrics. The consistently high F1-score across all models (especially for the N-Gram and combined models) indicates a good balance between precision and recall. This means that the models are not only accurate in their predictions (high precision) but also able to identify most of the relevant sentiments (high recall).

To better illustrate the performance comparison between the three Naïve Bayes models using different feature sets, Figure 3 presents a bar chart visualizing the key evaluation metrics. This chart provides a clear, side-by-side comparison of the Lexicon-based, N-Gram, and Combined Lexicon-N-Gram models across four critical performance indicators: Accuracy, Precision, Recall, and F1-score.

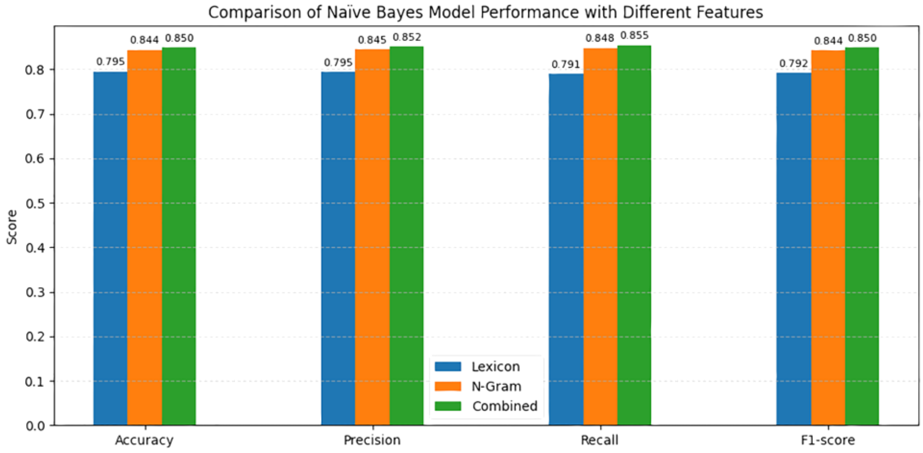

Fig. 4. Visualization of Model Comparison

As evident from the visualization, the Combined Lexicon-N-Gram model consistently outperforms the other two models across all metrics, albeit by a small margin in some cases. The N-Gram model shows significant improvement over the Lexicon-based model, while the Combined model further refines these gains. This visual representation reinforces our earlier discussion on the synergistic effects of combining Lexicon and N-Gram features, demonstrating how this approach leverages the strengths of both methods to achieve superior performance in sentiment classification tasks.

In the context of e-commerce sentiment analysis, the better performance of the N-Gram and combined models can be explained by their ability to capture phrases and word combinations that frequently appear in reviews. For example, phrases such as "very satisfied" or "poor quality" may be more informative than individual words in determining sentiment. The combined Lexicon-N-Gram model showed a slight improvement over the pure N-Gram model, indicating that the addition of lexicon information may help in cases where N-Gram context alone may not be sufficient. This can be especially useful in handling language variations and idiomatic expressions that may not be captured by N-Grams alone. Even though the Lexicon model shows lower performance, the accuracy and F1-score values above 77% still show the usefulness of the dictionary-based approach in sentiment analysis. This approach may be more suitable for cases where domain-specific vocabulary is critical or when the training dataset is limited. Overall, these results show that the combined approach of Lexicon-N-Gram with Naïve Bayes algorithm offers the most effective method for sentiment analysis in the context of Shopee e-commerce reviews. This improved performance is likely due to the model's ability to capture both semantic information (via lexicon) and contextual patterns (via N-grams) in the review text.

TABLE VIII
NAÏVE BAYES TRAINING RESULT AFTER CROSS VALIDATION

Method	Accuracy	Precision	Recall	F1-Score
Naïve Bayes Model with Lexicon Features	77%	78.8%	81%	80%
Naïve Bayes Model with N-Gram Features	82.8%	87.3%	80.4%	83.7%
Naïve Bayes Model with Combined Lexicon – N-Gram Features	83.4%	89.5%	79.1%	84%

A deeper analysis of the confusion matrices provides further insights into the models' performance. For the Lexicon model, out of 8,213 total predictions, there were 2,038 true negatives and 3,022 true positives, indicating good performance in correctly identifying both negative and positive sentiments. However, it also had 812 false positives and 698 false negatives, suggesting some difficulty in distinguishing between sentiments in certain cases. This resulted in an accuracy of 77%, with a slightly higher recall (81%) than precision (78.8%), indicating a tendency to slightly over-predict positive sentiments. The N-Gram model showed significant improvement, with 2,547 true negatives and 2,898 true positives, while reducing false positives to 420 and false negatives to 705. This led to a higher accuracy of 82.8%, with a notably improved precision of 87.3%, indicating fewer false positive predictions. However, the recall remained similar at 80.4%, suggesting that while the N-Gram model was more cautious in predicting positive sentiments, it didn't significantly improve in identifying all actual positive sentiments. The Combined Lexicon and N-Gram model further improved performance with 2,635 true negatives and 2,850 true positives, while having only 332 false positives but a slightly increased 753 false negatives. This resulted in the highest accuracy of 83.4% and the best precision of 89.5%, indicating very few false positive predictions. However, the recall decreased slightly to 79.1%, suggesting that while this model was the most accurate in its positive predictions, it might have been overly cautious, missing some actual positive sentiments. The F1-score of 84% for this model balances these trade-offs, confirming its overall superior performance. These results highlight the Combined model's strength in reducing false positives, which is particularly valuable in e-commerce contexts where incorrectly identifying negative sentiments as positive could lead to misguided business decisions.

The Naïve Bayes model with combined features demonstrates superior performance compared to the other two approaches due to several key aspects. Firstly, this model creates a synergy between semantic information from the lexicon and contextual information from N-Grams, enabling a deeper understanding of sentiment in the e-commerce context. Secondly, this combination is more effective in handling language variations and idiomatic expressions common in online reviews. Thirdly, the feature combination helps mitigate the weaknesses of each approach, resulting in more robust sentiment classification. The combined model also shows improvement in handling negation and intensification, as well as better adaptation to e-commerce-specific terminology. Furthermore, this combination enhances the model's generalization capability, allowing it to better handle review variations that might not be seen in the training data. Lastly, the combined model is more effective in dealing with ambiguity of words that may have different sentiments in different contexts. Although the accuracy improvement from 84.6% (N-Gram model) to 85% (combined model) seems small, it is significant in the context of e-commerce sentiment analysis, indicating the model's ability to handle more complex or ambiguous cases. On a large scale, this improvement can have a significant impact on customer sentiment analysis and business decision-making, potentially leading to more accurate customer insights and improved service strategies.

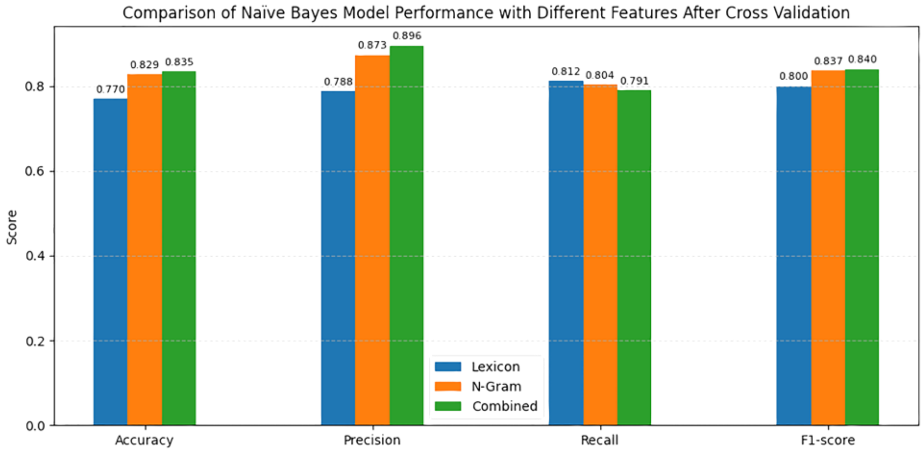

Fig. 5. Visualization of Model Comparison After Cross Validation

The findings of this study align with and extend previous research in the field of e-commerce sentiment analysis. Our results corroborate the findings of Abighail et al. [6], who achieved 72% accuracy using Naïve Bayes with TF-IDF weighting on Shopee reviews. Our improved accuracy of 85% with the combined Lexicon-N-Gram approach demonstrates the potential of feature combination in enhancing model performance. Similarly, our results surpass the 81.13% precision reported by Bahtiar et al. [7] for Naïve Bayes on marketplace reviews, highlighting the effectiveness of our combined feature approach. The high accuracy achieved in our study also aligns with the findings of Mantik et al. [8], who reported 94% accuracy using Naïve Bayes with WordCloud on online store reviews, although our study used a larger and more diverse dataset. These findings have significant implications for both theory and practice in e-commerce sentiment analysis. Theoretically, our results support the hypothesis that combining lexicon-based and N-gram features can lead to improved sentiment classification accuracy, suggesting a potential direction for future research in feature engineering for sentiment analysis. The superior performance of the combined model also underscores the importance of considering both semantic and contextual information in sentiment analysis tasks, particularly in the e-commerce domain where reviews often contain domain-specific terminology and complex sentiment expressions.

From a practical standpoint, the improved accuracy of our combined model could lead to more reliable sentiment analysis in e-commerce platforms. This could enhance customer feedback processing, allowing businesses to more accurately gauge customer satisfaction and identify areas for improvement. The ability to handle complex and ambiguous cases more effectively could also lead to more nuanced insights into customer sentiments, potentially informing product development, customer service strategies, and marketing efforts. However, it's important to note that while our results show improvement over some previous studies, they don't reach the 94% accuracy reported by Mantik et al. [8]. This discrepancy might be due to differences in dataset size, diversity, or specific characteristics of the reviews analyzed. Future research could explore these factors to further enhance the performance and generalizability of sentiment analysis models in e-commerce contexts.

In this study, the total amount of data used was 8213, which reflects the total number of entries in the dataset. When applying the cross-validation method, this process divides the dataset into multiple folds, where the model is trained and tested in turn. The results of each fold produce a confusion matrix that records the number of correct and incorrect positive and negative predictions. The total number of elements in all the confusion matrices of each scenario generated from each fold must equal the number of the initial dataset, which is 8213. This ensures that the model evaluation is thorough, and no data is missed, providing an accurate picture of the model's performance in sentiment classification. Thus, this analysis demonstrates integrity and consistency in data processing, as well as the validity of the results obtained from model testing.

The dataset size of 8,213 entries used in this study warrants further discussion regarding its impact on the results and the potential for generalization to a broader population. While this sample size provides substantial data for analysis and generally good statistical power, it's important to consider whether it truly represents the entire population of Shopee reviews. The representativeness depends on factors such as the time frame of data collection, the diversity of products reviewed, and the demographic spread of reviewers.

The performance of the Naive Bayes models on this dataset, with accuracies ranging from 77.3% to 85.2%, suggests that the size is sufficient for effective model training and evaluation. However, it's crucial to consider how changes in dataset size might affect the results. A larger dataset could potentially improve model performance due to more training examples, provide greater confidence in the generalizability of results, and possibly capture more diverse language patterns and sentiments. Conversely, a smaller dataset might lead to overfitting, less reliable performance metrics due to increased variance, and a reduced ability to capture the full spectrum of sentiment expressions. While the current results provide valuable insights, caution should be exercised in generalizing these findings to all e-commerce reviews or even all Shopee reviews. The specific context of the collected data may influence the sentiment patterns observed. Moreover, the relative performance of lexicon-based, N-gram, and combined features might shift with changes in dataset size. For instance, N-gram features might become more effective with larger datasets as they capture more language patterns. To enhance the robustness and generalizability of these findings, future research could consider expanding the dataset size to test if the observed patterns hold, conducting a sensitivity analysis with varying dataset sizes to understand how model performance scales, and incorporating data from different time periods or product categories to ensure broader representativeness. While the current dataset provides valuable insights, acknowledging these limitations and potential impacts on results is crucial for a comprehensive understanding of the study's implications and its place within the broader context of e-commerce sentiment analysis.

Regarding the generalizability of these research findings to other contexts, several important considerations arise. While this study focused on the Shopee platform, the methodology employed - particularly the combination of Lexicon-N-Gram features with Naive Bayes - potentially has applicability to other e-commerce platforms such as Tokopedia, Lazada, or even Amazon. However, its effectiveness may vary depending on the specific characteristics of reviews on these platforms, including differences in language, writing styles, or platform-specific terminology. In terms of application to other product categories, the model is likely to adapt well, especially for products with review characteristics similar to those in the dataset used in this study. However, for highly specific or technical product categories, adjustments to the lexicon dictionary or model retraining might be necessary to accommodate specialized terminology. It's important to note that these generalizations require further empirical validation. Future research could focus on applying this model across various e-commerce platforms and product categories to more comprehensively test the robustness and adaptability of this method.

IV. Conclusion

This research investigated sentiment classification in e-commerce using Naïve Bayes with lexicon-based, N-gram, and combined lexicon-N-gram features. The study analyzed customer reviews of the Shopee application from the Google Play Store. Our findings demonstrate that the Naïve Bayes model using combined lexicon-N-gram features achieved the highest performance among the three methods, with an accuracy of 84.9%, precision of 85.2%, recall of 85.4%, and an F1-score of 84.9%. This superior performance can be attributed to the combination of lexicon and N-gram features capturing a wider range of sentiment indicators, leading to improved sentiment identification.

The N-gram model also performed well, with an accuracy of 84.4%, precision of 84.5%, recall of 84.8%, and an F1-score of 84.3%. This suggests that N-grams effectively capture contextual information from the text, contributing to accurate sentiment classification. The lexicon-based model, while still effective, showed slightly lower metrics with an accuracy of 77.9%. This indicates that predefined sentiment dictionaries may not capture contextual nuances as effectively as N-grams or the combined approach.

Our research demonstrates notable improvements in model performance compared to some previous studies in the field. The final evaluation metrics for the Naïve Bayes Model with Combined Lexicon-N-gram Features, including an accuracy of 83.4% and a precision of 89.5%, highlight the effectiveness of this approach in sentiment classification. These findings contribute to the field of sentiment analysis in e-commerce, offering insights into the effectiveness of different feature extraction methods when used with Naïve Bayes classifiers. The study underscores the importance of contextual information in sentiment analysis and provides a foundation for further research in improving sentiment classification techniques for e-commerce platforms. Future work could explore more advanced feature combination techniques or investigate the performance of these models on larger and more diverse datasets.

References

[1] V. Bonta, N. Kumaresh, and N. Janardhan, “A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis,” *Asian Journal of Computer Science and Technology*, vol. 8, no. S2, pp. 1–6, Mar. 2019, doi: 10.51983/ajcst-2019.8.s2.2037.

[2] G. Aliman *et al.*, “Sentiment Analysis using Logistic Regression,” 2022.

[3] B. Gunawan, H. S. Pratiwi, and E. E. Pratama, “Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes,” vol. 4, no. 2, pp. 17–29, 2018.

[4] P. A. Permatasari, L. Linawati, and L. Jasa, “Survei Tentang Analisis Sentimen Pada Media Sosial,” *Majalah Ilmiah Teknologi Elektro*, vol. 20, no. 2, p. 177, Dec. 2021, doi: 10.24843/mite.2021.v20i02.p01.

[5] J. M. Bisnis, D. Saing, and U. Alwendi, “Penerapan E-Commerce Dalam Meningkatkan,” vol. 17, no. 3, 2020, [Online]. Available: http://jurnal.undiknas.ac.id/index.php/magister-manajemen/

[6] B. M. D. Abighail, Fachrifansyah, M. R. Firmanda, M. S. Anggreainy, Harvianto, and Gintoro, “Sentiment Analysis E-commerce Review,” in *Procedia Computer Science*, Elsevier B.V., 2023, pp. 1039–1045. doi: 10.1016/j.procs.2023.10.613.

[7] S. A. H. Bahtiar, C. K. Dewa, and A. Luthfi, “Comparison of Naïve Bayes and Logistic Regression in Sentiment Analysis on Marketplace Reviews Using Rating-Based Labeling,” *Journal of Information Systems and Informatics*, vol. 5, no. 3, pp. 915–927, Aug. 2023, doi: 10.51519/journalisi.v5i3.539.

[8] J. Mantik, E. R. Putri, and H. Februariyanti, “Product Review Sentiment Analysis At Online Store Jiniso Official Shop Using Naive Bayes Classifier (Nbc) Method,” Online, 2022.

[9] S. Dey, S. Wasif, D. S. Tonmoy, S. Sultana, J. Sarkar, and M. Dey, “A Comparative Study of Support Vector Machine and Naive Bayes Classifier for Sentiment Analysis on Amazon Product Reviews,” in *2020 International Conference on Contemporary Computing and Applications, IC3A 2020*, Institute of Electrical and Electronics Engineers Inc., Feb. 2020, pp. 217–220. doi: 10.1109/IC3A48958.2020.233300.

[10] C. Juliane, “Implementation of Naive Bayes Algorithm on Sentiment Analysis Application,” 2021.

[11] S. Elbagir and J. Yang, *Twitter Sentiment Analysis Using Natural Language Toolkit and VADER sentiment*. 2019.

[12] Y. Qi and Z. Shabrina, “Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach,” *Soc Netw Anal Min*, vol. 13, no. 1, Dec. 2023, doi: 10.1007/s13278-023-01030-x.

[13] R. Mahendrajaya, G. A. Buntoro, and M. B. Setyawan, “ANALISIS SENTIMEN PENGGUNA GOPAY MENGGUNAKAN METODE LEXICON BASED DAN SUPPORT VECTOR MACHINE,” 2019. [Online]. Available: http://studentjournal.umpo.ac.id/index.php/komputek

[14] O. Manullang, C. Prianto, and N. H. Harani, “Analisis Sentimen Untuk Memprediksi Hasil Calon Pemilu Presiden Menggunakan Lexicon Based dan Random Forest,” 2023.

[15] P. Pratama, E. Indarbensyah, and N. Rochmawati, “Penerapan N-Gram menggunakan Algoritma Random Forest dan Naïve Bayes Classifier pada Analisis Sentimen Kebijakan PPKM 2021,” *Journal of Informatics and Computer Science*, vol. 02, 2021.

[16] A. M. Priyatno and F. I. Firmananda, “N-Gram Feature for Comparison of Machine Learning Methods on Sentiment in Financial News Headlines,” *RIGGS: Journal of Artificial Intelligence and Digital Business*, vol. 1, no. 1, pp. 01–06, Jul. 2022, doi: 10.31004/riggs.v1i1.4.

[17] N. Ika, P. Kalingara, O. N. Pratiwi, and H. D. Anggana, “ANALISIS SENTIMEN REVIEW CUSTOMER TERHADAP LAYANAN EKSPEDISI JNE DAN J&T EXPRESS MENGGUNAKAN METODE NAÏVE BAYES SENTIMENT ANALYSIS REVIEW CUSTOMER OF JNE AND J&T EXPRESS EXPEDITION SERVICES USING NAÏVE BAYES METHOD,” vol. 8, no. 5, 2021.

[18] M. R. Romadhon and F. Kurniawan, “A Comparison of Naive Bayes Methods, Logistic Regression and KNN for Predicting Healing of Covid-19 Patients in Indonesia,” in *3rd 2021 East Indonesia Conference on Computer and Information Technology, EIConCIT 2021*, Institute of Electrical and Electronics Engineers Inc., Apr. 2021, pp. 41–44. doi: 10.1109/EIConCIT50028.2021.9431845.

[19] A. Guswandri, R. P. Cahyono, S. I. Akutansi, and T. Komputer, “PENERAPAN SENTIMEN ANALIS MENGGUNAKAN METODE NAÏVE BAYES DAN SVM,” 2022.

[20] A. Sabrani, I. W. Gede Putu Wirarama Wedashwara, and F. Bimantoro, “METODE MULTINOMIAL NAÏVE BAYES UNTUK KLASIFIKASI ARTIKEL ONLINE TENTANG GEMPA DI INDONESIA,” 2020. [Online]. Available: http://jtika.if.unram.ac.id/index.php/JTIKA/