LEXICON-BASED AND NAIVE BAYES SENTIMENT ANALYSIS FOR RECOMMENDING THE BEST MARKETPLACE SELECTION AS A MARKETING STRATEGY FOR MSMES

Hoiriyah[1]; Helva Mardiana[2]; Miftahul Walid[3]; Aang Kisnu Darmawan[4]

[1,4] Department of Information System, [3,4]Department of Informatics Engineering
[1234]Universitas Islam Madura
https://www.uim.ac.id/
[*1]hoiriyah.file.uim@gmail.com, [2]helvadiana26@gmail.com,[3] miftahul.walid@uim.ac.id,
[4]ak.darmawan@gmail.com

(*) Corresponding Author

Abstract—MSMEs (micro, small, and medium enterprises) play an essential role in the Indonesian economy, contributing to 60% of the country's GDP (gross domestic product), creating jobs, and increasing non-oil and gas exports. However, MSMEs in Indonesia face various challenges, including access to technology, digital marketing tools, financial resources, limited market distribution, and low technological literacy. Marketplaces provide an essential marketing channel for MSMEs to increase their competitiveness and sales. Sentiment analysis can assist businesses in making informed decisions about which marketplace to use to increase customer satisfaction. Apart from the importance of the marketplace for MSMEs in Indonesia, research on sentiment analysis for marketplace recommendations is still minimal. Therefore, this study aims to analyze six popular marketplaces in Indonesia using Lexicon-based and naïve Bayes research methods to provide the best marketplace recommendations for MSME marketing. The results showed that Blibli.com had the highest accuracy, followed by Tokopedia, Tiktokshop, Lazada, Shopee, and Bukalapak. Blibli.com received positive reviews with 96.33%, followed by Tokopedia with 95.25%, Tiktokshop with 94.61%, and Lazada with the highest accuracy. 94.22%, Shopee 92.18%, and Bukalapak 89.57%. This research has two significant contributions. First, making a scientific contribution by applying a combination model of lexicon-based and naïve Bayes to analyze market sentiment in Indonesia Second, offering a practical contribution by providing recommendations to MSME actors and policymakers in choosing the best marketplace for MSMEs marketing purposes in Indonesia. By utilizing the recommended marketplace, MSMEs can optimize their marketing strategy and increase their competitiveness in the digital marketplace.

Keywords: lexicon-based, marketplace, MSMEs, naïve Bayes, sentiment analysis.

Abstrak—*UMKM (Usaha Mikro, Kecil dan Menengah) memainkan peran penting dalam perekonomian Indonesia, berkontribusi terhadap 60% PDB (Produk Domestik Bruto) negara, menciptakan lapangan kerja, dan meningkatkan ekspor nonmigas. Namun, UMKM di Indonesia menghadapi berbagai tantangan, antara lain akses teknologi, alat pemasaran digital, sumber daya keuangan dan distribusi pasar yang terbatas serta literasi teknologi yang tidak memada., Marketplace menyediakan saluran pemasaran penting bagi UMKM untuk meningkatkan daya saing dan meningkatkan penjualan. Analisis sentimen dapat membantu bisnis dalam membuat keputusan berdasarkan informasi tentang marketplace mana yang akan digunakan untuk meningkatkan kepuasan pelanggan. Terlepas dari pentingnya marketplace bagi UMKM di Indonesia, penelitian tentang analisis sentimen untuk rekomendasi marketplace masih sangat minim. Oleh karena itu, penelitian ini bertujuan untuk menganalisis enam marketplace populer di Indonesia dengan menggunakan metode penelitian Lexicon-based dan naïve Bayes untuk memberikan rekomendasi marketplace terbaik untuk pemasaran UMKM. Hasil penelitian menunjukkan bahwa Blibli.com memiliki akurasi tertinggi, diikuti oleh Tokopedia, Tiktokshop, Lazada, Shopee, dan Bukalapak. Blibli.com memperoleh review positif dengan 96,33%, diikuti oleh Tokopedia dengan 95,25%, Tiktokshop dengan 94,61%, Lazada dengan akurasi tertinggi. 94,22%, Shopee 92,18%, dan Bukalapak 89,57%. Penelitian ini memiliki dua kontribusi signifikan. Pertama, memberikan kontribusi ilmiah dengan menerapkan model kombinasi lexicon-based Dan naïve Bayes untuk analisis sentimen pasar di Indonesia. Kedua, menawarkan kontribusi praktis dengan memberikan rekomendasi kepada pelaku UMKM dan pembuat kebijakan dalam memilih marketplace terbaik untuk tujuan pemasaran UMKM di Indonesia. Dengan memanfaatkan marketplace yang direkomendasikan, UMKM dapat mengoptimalkan*

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 19, No. 2 September 2023 | DOI: 10.33480/pilar.v19i2.4176

strategi pemasarannya dan meningkatkan daya saingnya di pasar digital.

***Kata Kunci**: lexicon-based, marketplace, umkm, naïve bayes, analisis sentimen.*

INTRODUCTION

One type of business with the most significant influence in the country's economic sector is the Micro, Small, and Medium Enterprises (MSMEs). Based on the data obtained, the number of MSMEs in Indonesia has continued to grow yearly, as recorded by the Ministry of Cooperatives and SMEs on their official website. Dataindonesia.id by (Mahdi, 2022) Based on data, in 2019, the development of MSMEs in Indonesia increased by 1.98% compared to the previous year. This indicates that MSMEs in Indonesia now account for 99.99% of the total number of businesses in the country. MSMEs play a crucial role in supporting the Indonesian economy, especially in light of the economic recession that Indonesia is projected to face in 2023. According to data *databooks.katadata.co.id* (Ahdiat, 2022) In 2021, MSMEs in Indonesia could absorb around 97% of the workforce in the country, contributing to 60.3% of the GDP and 14.4% of the national sector. Based on this data, the President directed that a digital marketing strategy be implemented to improve the performance of MSMEs nationwide. This was reported by (Doni, 2021). On Thursday, June 10, 2021, President Joko Widodo announced during a meeting at the Merdeka Palace that the government would strive to digitize the sales of MSME products, with a target of 30 million MSMEs entering the digital ecosystem by 2024.

The MSMEs sector in Indonesia is currently facing various challenges, particularly in marketing. These challenges have been exacerbated by the COVID-19 pandemic, with approximately 63.9% of MSMEs experiencing a decrease in revenue by more than 30%. However, in contrast to this trend, the online marketplace or e-commerce has seen increased online transactions during the pandemic. According to the Director-General of Post and Informatics Administration (PPI) at the Ministry of Communication and Information Technology, the number of online transactions during the pandemic has increased significantly by up to 400%. (Pamungkas, 2023). According to data for (Kemp, 2021), In January 2021, there were approximately 175.4 million internet users in Indonesia. Around 160.0 million internet users (91.2%) accessed the internet through mobile devices. The number of e-commerce users in Indonesia is estimated to reach 74.7 million in 2021. The e-commerce market in Indonesia continues to grow, driven by the increasing number of internet users and the widespread penetration of smartphones. There are many marketplace platforms in Indonesia, among which the popular ones are Shopee, Tokopedia, Lazada, Blibli.com, Bukalapak, and even the latest one, TikTokShop, which can now compete with other platforms. However, the choice of the marketplace can impact product sales strategy. As the most appropriate marketing strategy, owners of MSMEs should determine relevant application platforms that align with the target market. Therefore, to realize the President's efforts in digitizing the marketing of MSMEs products, one of the efforts is to conduct sentiment analysis on the best and most relevant e-marketplaces following user sentiment as a recommendation for MSMEs owners to implement targeted marketing strategies.

In 2022, a similar study was conducted by Tito Dwiki Darmawan, which focused on sentiment analysis of customer reviews in Indonesian e-commerce. The study was published in a paper titled "Analisis Sentimen Review Pelanggan *E-commerce* di Indonesia Menggunakan Algoritma Naïve Bayes." This research utilized the Naive Bayes classification method to determine the public sentiment toward e-commerce, resulting in positive sentiment towards Lazada at 97.0% with an accuracy of 56.23%. In the second position was Bukalapak, with 94.6% positive sentiment and an accuracy of 93.0%, followed by Shopee at 88.5% with an accuracy of 87.82%. Blibli.com followed with a positive sentiment of 76.1% and an accuracy of 55.31%. Lastly, Tokopedia had a positive sentiment of 34.4% with an accuracy of 94.94%. (Darmawan, 2022). Furthermore, in 2021, Dwi Latifah Rianti et al. conducted a study on the marketplace trend based on customer review classification using kernel comparison in the support vector machine method. The study aimed to classify the most relevant marketplaces for selling their products. This was done because business actors did not know enough about how the choice of a marketplace could affect their product sales strategy. Therefore, this research was conducted to support business actors in seeing the community's tendencies in choosing a marketplace. Based on the kernel model comparison in the SVM method, the results showed that the sigmoid kernel model was the most suitable kernel model for the classification process in this study, with accuracy, precision, Recall, and F1 score of 92% and parameters of C=100, γ=0.01, and r=1. From the results obtained, it can be concluded that the marketplace trend based on the highest reviews is ranked Tokopedia, followed by Shopee, and lastly, Bukalapak. (Rianti, Umaidah, & Voutama, 2021). Furthermore, in 2022, Destaria Wilandini and Purwantoro conducted a study on social media to observe culinary trends

using the naive Bayes method. The study results indicated that TikTok is the most frequently used application for observing culinary trends, thus making it the most recommended choice. Following TikTok were Instagram, Twitter, YouTube, and finally Facebook. (Wilandini & Purwantoro, 2022), The study only focused on one object, which is culinary.

Based on the literature from previous research, there are several differences in research with previous researchers, including the source of data that will be obtained through crawling Twitter data using Netlytic tools. In addition, six e-marketplace platforms will be analyzed in this study, namely Shopee, Lazada, Tokopedia, Tiktkshop, blibli.com, and Bukalapak, according to the development of the marketplace each year. So far, there has never been any research on the Tiktokshop marketplace, which is proliferating. Moreover, this research to be conducted is a development of previous research. The upcoming research will provide a new comparative perspective by using different methods. In this research, lexicon-based and naive Bayes methods will be used. The lexicon-based method determines whether the words used in consumer reviews have a positive or negative sentiment toward a marketplace (Asri, Suliyanti, Kuswardani, & Fajri, 2022), while naive Bayes classifies consumer reviews as positive or negative. The naive Bayes method is one of the classification methods in Machine Learning. Naive Bayes is also a simple probabilistic prediction technique based on applying Bayes' theorem or rules, assuming strong independence among features. (Utama et al., 2019) This method has been frequently used in other studies with reasonable accuracy. (Hartatik, Tamam, & Setyanto, 2020) The Naive Bayes method also has excellent precision despite having less training data. (Umbu et al., 2022)

Furthermore, according to Larose in his book "Data Mining Methods and Models," data mining is extracting patterns (pattern recognition) from data in a database to gain knowledge. Meanwhile, the classification method is a data mining technique that can create a model to differentiate a concept or class whose labels are unknown. It can also be interpreted that data mining functions to classify data into specific classes based on categories and is a type of supervised learning. (Risnasari, 2022). In this research, the programming language used is Python because Python is considered one of the most straightforward programming languages. (Tamam et al., 2023)

Therefore, this research aims to determine public sentiment towards e-marketplace platforms and to investigate whether lexicon-based and naive Bayes methods can be used for sentiment weighting and classification with the most appropriate level of accuracy. It is hoped that this study will assist SMEs in developing their sales strategies by providing recommendations for marketplace platforms with the highest positive sentiment and the best accuracy level.

MATERIALS AND METHODS

The research methodology is a framework used by researchers to conduct a study. (Wirma, 2022) This phase includes detailed and systematic research steps, from data collection to analysis. The overview of the research stages is presented in a diagram format as follows:

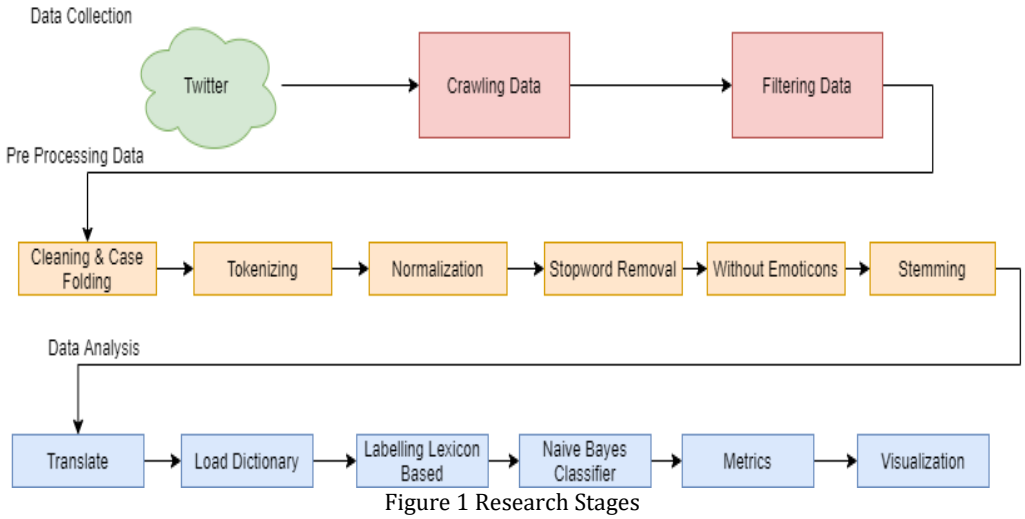


Figure 1 Research Stages

Below is an explanation of the stages of the diagram according to Figure 1.

Data Collection

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 19, No. 2 September 2023 | DOI: 10.33480/pilar.v19i2.4176

In the next stage, the data collection process will be carried out, where the author will collect sentiment data by crawling Twitter using the Netlytic tool (Johnson & Smith, 2022) at the following link https://netlytic.org/index.php. The following is the process of crawling data related to public sentiment towards e-marketplace platforms (Shopee, Tokopedia, Lazada, Tiktokshop, Bukalapak, and Blibli) using the keywords #shopee, #tokopedia, #lazada, #tiktokshop, #bukalapak, and #blibli. The data collected will only focus on Indonesian language data. The obtained data will be displayed and stored in a CSV format to facilitate sentiment analysis.

Data Pre Processing

Data pre-processing is one of the initial stages performed before conducting classification (Haranto & Sari, 2019). In this stage, the data obtained from the crawling process will be processed using the lexicon-based and naïve Bayes classifier methods. The data will be processed starting with data processing (data pre-processing).

In the pre-processing data stage, several steps are taken: *cleaning & case folding, tokenizing, normalization, stopword removal, removing emoticons, and stemming*. (Hasugian, Fakhriza, & Zukhoiriyah, 2023)

Data Analysis

In the analysis stage, the author attempted to analyze the processed data using Python to analyze the community's sentiment toward e-marketplace platforms. The sentiment weighting was done using the lexicon-based method by determining positive and negative sentiments obtained from the lexicon dictionary. Next, a classification process was carried out using the naive Bayes method to determine positive and negative sentiments from the labelling done through the previous weighting results. This process used the Python programming language (Mardiana, Syahreva, & Tuslaela, 2019) to obtain accurate results. The accuracy value is the percentage of the accuracy of a data record that will be classified correctly after testing the classification results.

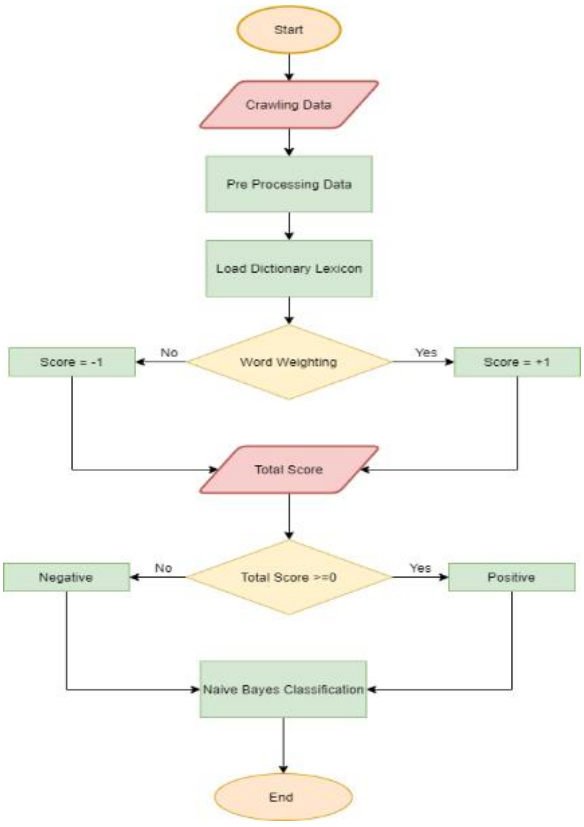


Figure 2 Research Stages

In Figure 2 above, there are several stages of the process. In this process, weighting is carried out using the lexicon-based method. In the data processing stage, the data is first crawled from social media Twitter. Then, a data pre-processing or data cleaning process is carried out with various steps, including cleansing, case folding, normalization, stopword removal, and tokenizing. After this point, Indonesian was translated into English using the Vader Sentiment library to read the lexicon dictionary. (Amaliah & Nuryana, 2022) Next, the process of loading the lexicon dictionary is carried out, which determines the sentiment of words using the lexicon dictionary. (Pratiwi & Nudin, 2021) Then, a weighting process is performed using the lexicon-based method, where if a word has a positive sentiment, it will have a value of +1, but if it has a negative sentiment, it will have a value of -1. (Kurniawan, Adinugroho, & Features, 2019) After obtaining the output, the resulting weighting is labelled as positive or negative, which can then undergo a classification process using the naïve Bayes classifier method. After classification, the testing process for confusion metrics and classification report (Faesal, Muslim, & Ruger, 2020) is performed to obtain the accuracy, precision, Recall, and F1 score values.

RESULTS AND DISCUSSION

In this chapter, we will discuss the results of data analysis based on data obtained from Twitter social media that has been crawled using Netlytic tools. The data contains comments or responses from the public on Twitter towards several popular marketplace applications in Indonesia, including Tiktokshop, Shopee, Lazada, Tokopedia, Bukalapak, and Blibli.com.

Data Collection Results

In this step, data crawling or data collection is carried out, in which the Netlytic tool is used on the following website. https://netlytic.org/index.php? For the data collection or Twitter crawling process, the tool is Netlytic with the following webpage and keywords: #shopee, #tokopedia, #lazada, #tiktokshop, #bukalapak, and #blibli. The data collected includes update (tweet date), author (username), and text (tweet). An example can be seen in the image below.

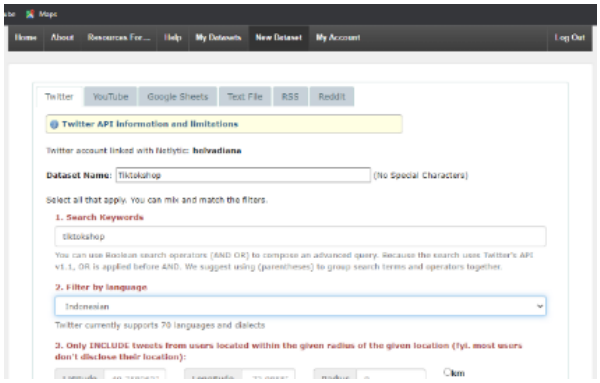


Figure 3 Tweet Data Crawling Process

Figure 3 above shows some datasets used to crawl data from Twitter according to the keywords. In the example image above, the dataset is named Tiktokshop with the keyword "@tiktokshop," and the language is Indonesian. In this process, Twitter data was taken in March 2023, with the acquisition of each data set as follows:

Table 1. Crawling Data Results

Marketplace	Crawling Result
Tiktokshop	2772
Shopee	9996
Lazada	4998
Blibli.com	4997
Bukalapak	4076
Tokopedia	4979

After the tweets data is collected, the obtained data is converted into Table 1 to format for easy processing in the next stage. The data frame contains three attributes, namely:

1. update: the time when the tweet was created
2. author: the username of the tweet maker
3. text: the text containing the content of the tweet that has been created.

Data Pre-Processing Results

1. Cleaning & Case Folding Result

In this stage, the raw data underwent several steps, including removing Twitter usernames, removing symbols, removing URL links (https and http), replacing HTML characters with quotation marks, removing punctuation, considering letters and numbers, replacing line breaks with spaces, converting every sentence to lowercase, removing single characters, and separating and merging data. Then, in this stage, duplicate and empty tweet data were removed. The following is the result of the cleaning and case folding process:

Table 2 Cleaning & Case Folding Results

Text	Clean
Aku beli kat tiktok shop and shopee serentak ye! Jgn tanya kenapa, aku pun taktau. Tiktokshop dah sampai, anyone yang nak beli shades 03, beli kat i!! I sini https://t.co/4dXn8RcXCd Harga skrg RM59 kat shopee, Harga aku beli RM49.60. Tak masuk delivery. Kalau berminat dm tauu! https://t.co/e8HhcaBadM	aku beli kat tiktok shop and shopee serentak ye jgn tanya kenapa aku pun taktau tiktokshop dah sampai anyone yang nak beli shades beli kat i i sini harga skrg rm kat shopee harga aku beli rm tak masuk delivery kalau berminat dm tauu
akutuu bingung pake tiktokshop. kenapa ongkir ke rumahku lebih mahal dripda shopee padahal barangnya mureh2 :(	akutuu bingung pake tiktokshop kenapa ongkir ke rumahku lebih mahal dripda shopee padahal barangnya mureh :(
Kenapa gue sangat norak, nyoba belanja online di tiktokshop aja bingungÃ°ÂŸÂ˜ÂÃ°ÂŸÂ˜Â	kenapa gue sangat norak nyoba belanja online di tiktokshop aja bingung

Table 2 above shows that there have been some changes to the results of the original data after the cleaning and case folding processes. In this cleaning process, the number of data can also be changed by removing duplicates and columns.

2. Tokenizing Result

In this step, the Python library NLTK is utilized to tokenize the text by breaking down a sentence into words or tokens, which can then be processed quickly. The following is the result of the tokenization process.

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 19, No. 2 September 2023 | DOI: 10.33480/pilar.v19i2.4176

Table 3 Tokenizing Results

Clean	Token
aku beli kat tiktok shop and shopee serentak ye jgn tanya kenapa aku pun taktau tiktokshop dah sampai anyone yang nak beli shades beli kat i i sini harga skrg rm kat shopee harga aku beli rm tak masuk delivery kalau berminat dm tauu	['aku', 'beli', 'kat', 'tiktok', 'shop', 'and', 'shopee', 'serentak', 'ye', 'jgn', 'tanya', 'kenapa', 'aku', 'pun', 'taktau', 'tiktokshop', 'dah', 'sampai', 'anyone', 'yang', 'nak', 'beli', 'shades', 'beli', 'kat', 'i', 'i', 'sini', 'harga', 'skrg', 'rm', 'kat', 'shopee', 'harga', 'aku', 'beli', 'rm', 'tak', 'masuk', 'delivery', 'kalau', 'berminat', 'dm', 'tauu']
akutuu bingung pake tiktokshop kenapa ongkir ke rumahku lebih mahal dripda shopee padahal barangnya mureh :(	['akutuu', 'bingung', 'pake', 'tiktokshop', 'kenapa', 'ongkir', 'ke', 'rumahku', 'lebih', 'mahal', 'dripda', 'shopee', 'padahal', 'barangnya', 'mureh', ':(']
kenapa gue sangat norak nyoba belanja online di tiktokshop aja bingung	['kenapa', 'gue', 'sangat', 'norak', 'nyoba', 'belanja', 'online', 'di', 'tiktokshop', 'aja', 'bingung']

Tokenizing is carried out in Table 3 above to make it easier to weigh the score in sentiment analysis.

3. Normalization Result

In this step, a process is carried out to convert a word to its correct form according to the KBBI (Great Dictionary of Indonesian Language) and make it understandable. Abbreviated terms such as 'yg' are changed to 'yang', 'tdk' to 'tidak', and so on by utilizing a normalization dictionary created before. The following is an example of tweet data after the normalization process.

Table 4 Normalization Results

Token	Normalization
['aku', 'beli', 'kat', 'tiktok', 'shop', 'and', 'shopee', 'serentak', 'ye', 'jgn', 'tanya', 'kenapa', 'aku', 'pun', 'taktau', 'tiktokshop', 'dah', 'sampai', 'anyone', 'yang', 'nak', 'beli', 'shades', 'beli', 'kat', 'i', 'i', 'sini', 'harga', 'skrg', 'rm', 'kat', 'shopee', 'harga', 'aku', 'beli', 'rm', 'tak', 'masuk', 'delivery', 'kalau', 'berminat', 'dm', 'tauu']	['aku', 'beli', 'di', 'tiktok', 'shop', 'and', 'shopee', 'bareng', 'ye', 'jangan', 'tanya', 'kenapa', 'aku', 'juga', 'tidak tahu', 'tiktokshop', 'sudah', 'sampai', 'anyone', 'yang', 'mau', 'beli', 'shades', 'beli', 'di', 'aku', 'aku', 'sini', 'harga', 'sekarang', 'ringgit', 'di', 'shopee', 'harga', 'aku', 'beli', 'ringgit', 'tidak', 'masuk', 'delivery', 'kalau', 'berminat', 'direct message', 'tau']
['akutuu', 'bingung', 'pake', 'tiktokshop', 'kenapa', 'ongkir', 'ke', 'rumahku', 'lebih', 'mahal', 'dripda', 'shopee', 'padahal', 'barangnya', 'mureh', ':(']	['aku', 'itu', 'bingung', 'pakai', 'tiktokshop', 'kenapa', 'ongkos kirim', 'ke', 'rumahku', 'lebih', 'mahal', 'dari pada', 'shopee', 'padahal', 'barangnya', 'murah', ':(']
['kenapa', 'gue', 'sangat', 'norak', 'nyoba', 'belanja', 'online', 'di', 'tiktokshop', 'aja', 'bingung']	['kenapa', 'aku', 'sangat', 'norak', 'nyoba', 'belanja', 'online', 'di', 'tiktokshop', 'saja', 'bingung']

After the normalization process in Table 4, several words have changed from the tokenization process. This can help simplify the sentiment analysis process for weighting meanings.

4. Stopword Removal Result

In this process, stopwords or irrelevant words in the Indonesian language are removed from the text. First, the NLTK (Natural Language Toolkit) module is imported, and the list of Indonesian stopwords from NLTK is downloaded. Then, the list of Indonesian stopwords from the manually created CSV file is loaded and saved in CSV format. In this process, the list of stopwords in the CSV file is converted into a set, and some specific stopwords such as "yg", "utk", "tdk", or some meaningless words like "adalah," "tidk", "jangan," etc., are also added to the "stopwords" variable. Here is an example of the Stopword Removal result.

Table 5 Stopword Removal Results

Normalization	Stopword Removal
['aku', 'beli', 'di', 'tiktok', 'shop', 'and', 'shopee', 'bareng', 'ye', 'jangan', 'tanya', 'kenapa', 'aku', 'juga', 'tidak tahu', 'tiktokshop', 'sudah', 'sampai', 'anyone', 'yang', 'mau', 'beli', 'shades', 'beli', 'di', 'aku', 'aku', 'sini', 'harga', 'sekarang', 'ringgit', 'di', 'shopee', 'harga', 'aku', 'beli', 'ringgit', 'tidak', 'masuk', 'delivery', 'kalau', 'berminat', 'direct message', 'tau']	['beli', 'tiktok', 'shop', 'and', 'shopee', 'bareng', 'tidak tahu', 'tiktokshop', 'anyone', 'beli', 'shades', 'beli', 'harga', 'ringgit', 'shopee', 'harga', 'beli', 'ringgit', 'masuk', 'delivery', 'berminat', 'direct message', 'tahu']
['aku', 'itu', 'bingung', 'pakai', 'tiktokshop', 'kenapa', 'ongkos kirim', 'ke', 'rumahku', 'lebih', 'mahal', 'dari pada', 'shopee', 'padahal', 'barangnya', 'murah', ':(']	['bingung', 'pakai', 'tiktokshop', 'ongkos kirim', 'rumahku', 'mahal', 'dari pada', 'shopee', 'barangnya', 'murah', ':(']
['kenapa', 'aku', 'sangat', 'norak', 'nyoba', 'belanja', 'online', 'di', 'tiktokshop', 'saja', 'bingung']	['norak', 'nyoba', 'belanja', 'online', 'tiktokshop', 'bingung']

Table 5 above shows the results of the stopword removal method, which removes words like "aku," "di", and "iya" and other associated expressions that are believed to be worthless.

5. Remove Emoticons Result

In this process, data cleaning is performed on several emoticon symbols that are considered to make sentiment analysis difficult. Therefore, these emoticons must be removed to facilitate the sentiment analysis process with the best results. Here is an example of the results of the emoticon removal process.

Table 6 Remove Emoticons Results

Stopword Removal	Remove Emoticons
['beli', 'tiktok', 'shop', 'and', 'shopee', 'bareng',	['beli', 'tiktok', 'shop', 'and', 'shopee', 'bareng',

Stopword Removal	Remove Emoticons
'tidak tahu', 'tiktokshop', 'anyone', 'beli', 'shades', 'beli', 'harga', 'ringgit', 'shopee', 'harga', 'beli', 'ringgit', 'masuk', 'delivery', 'berminat', 'direct message', 'tahu']	'tidak tahu', 'tiktokshop', 'anyone', 'beli', 'shades', 'beli', 'harga', 'ringgit', 'shopee', 'harga', 'beli', 'ringgit', 'masuk', 'delivery', 'berminat', 'direct message', 'tahu']
['bingung', 'pakai', 'tiktokshop', 'ongkos kirim', 'rumahku', 'mahal', 'dari pada', 'shopee', 'barangnya', 'murah', ':(']	['bingung', 'pakai', 'tiktokshop', 'ongkos kirim', 'rumahku', 'mahal', 'dari pada', 'shopee', 'barangnya', 'murah']
['norak', 'nyoba', 'belanja', 'online', 'tiktokshop', 'bingung']	['norak', 'nyoba', 'belanja', 'online', 'tiktokshop', 'bingung']

Table 6 shows that there are emoticons in the second point that are removed in the process of removing emoticons.

6. Stemming result

In this process, a stemming pre-processing step is performed by utilizing the stemmer factory library in the Python programming language. This is done to facilitate data processing. The purpose of this stemming step is to remove affixes and replace words with their base form. Here are the results of the steaming process.

Table 7 Stemming Results

Remove Emoticons	Stemming
['beli', 'tiktok', 'shop', 'and', 'shopee', 'bareng', 'tidak tahu', 'tiktokshop', 'anyone', 'beli', 'shades', 'beli', 'harga', 'ringgit', 'shopee', 'harga', 'beli', 'ringgit', 'masuk', 'delivery', 'berminat', 'direct message', 'tahu']	['beli', 'tiktok', 'shop', 'and', 'shopee', 'bareng', 'tidak tahu', 'tiktokshop', 'anyone', 'beli', 'shades', 'beli', 'harga', 'ringgit', 'shopee', 'harga', 'beli', 'ringgit', 'masuk', 'delivery', 'minat', 'direct message', 'tahu']
['bingung', 'pakai', 'tiktokshop', 'ongkos kirim', 'rumahku', 'mahal', 'dari pada', 'shopee', 'barangnya', 'murah']	['bingung', 'pakai', 'tiktokshop', 'ongkos kirim', 'rumah', 'mahal', 'dari pada', 'shopee', 'barang', 'murah']
['norak', 'nyoba', 'belanja', 'online', 'tiktokshop', 'bingung']	['norak', 'nyoba', 'belanja', 'online', 'tiktokshop', 'bingung']

The word "berminat" in the first table is converted to "minat" due to the steaming process, as seen in Table 7 above.

Data Analysis Results

1. Translate the result

After the tweets data has undergone a series of pre-processing steps, the next step is to translate the data into English using a translator library in the Python programming language. The goal is to facilitate the following process: to perform weighting and labelling using the VADER sentiment library, which uses English. Here is an example of the result of the data translation.

Table 8 Translate Results

Teks	Text
beli tiktok shop and shopee bareng tidak tahu tiktokshop anyone beli shades beli harga ringgit shopee harga beli ringgit masuk delivery minat direct message tahu	buy tiktok shop and shopee together don't know tiktokshop anyone buy shades buy ringgit price shopee buy ringgit price enter delivery interest direct message you know
bingung pakai tiktokshop ongkos kirim rumah mahal dari pada shopee barang murah	confused about using the Tiktok shop, the cost of sending home is expensive, rather than the cheap goods shop
norak nyoba belanja online tiktokshop bingung	tacky trying to shop online tiktokshop confused

Table 8 above shows the translation results using the Python programming language library.

2. Load Dictionary Result

In this stage, a Lexicon dictionary is read using the Vader Sentiment library in any Python language. The purpose is to process data through lexicon-based weighting and labelling.

3. Lexicon-Based Weighting Result

After the lexicon input step, the next step is to perform a Compound Score calculation for weighting using the Vader Sentiment library, previously installed in Python. In this step, the sentiment score of each tweet is calculated using the VADER (Valence Dictionary and Sentiment Reasoned) library, which has been trained on a particular corpus of data. The formula for calculating the sentiment score is based on the value of positive words (+1), negative words (-1), and neutral words (0). The following is the result of lexicon-based weighting.

Table 9 Lexicon-Based Weighting Results

Text	Compound Score
buy tiktok shop and shoppe together don't know tiktokshop anyone buy shades buy ringgit price shoppe buy ringgit price enter delivery interest direct message you know	0.4588
confused about using the Tiktok shop, the cost of sending home is expensive, rather than the cheap goods shop	-0.3182
tacky trying to shop online tiktokshop confused	-0.3182

The results of lexicon-based weighting are shown in Table 9.

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 19, No. 2 September 2023 | DOI: 10.33480/pilar.v19i2.4176

4. Lexicon-Based Labeling Result

The labeling in this study uses the lexicon-based method by utilizing the VaderSentiment library in the Python programming language. In this process, tweet data is classified based on the result of the compound score weighting. Sentiments with a compound score <= 0 are considered harmful, while compound scores >= 0 are considered positive. The following are the results of lexicon-based labeling, shown in Table 10.

Table 10 Lexicon-Based Labeling Results

Text	Compound Score	Sentiments
buy tiktok shop and shopee together don't know tiktokshop anyone buy shades buy ringgit price shopee buy ringgit price enter delivery interest direct message you know	0.4588	Positive
confused about using the Tiktok shop, the cost of sending home is expensive, rather than the cheap goods shop	-0.3182	Negative
tacky trying to shop online tiktokshop confused	-0.3182	Negative

Table 10 shows that the compound score of more than or equal to 0 is labeled as positive, while the one with a value of -0 is labeled as unfavorable.

5. Naïve Bayes Classifier Results

The naïve Bayes method is implemented to classify positive and negative sentiments from tweet data. After the data has undergone pre-processing, weighting, and labeling using the lexicon-based method, the results of weighting and labeling are then classified using the naïve Bayes model. In this process, the selection of training data and testing data is carried out on the tweet dataset. Firstly, tweet data is separated based on the positive and negative sentiment labels into two sets, namely *set_positif,* and *set_negatif.* Next, the data in set_positif and set_negatif are randomly selected by 80% as the training data. This training data is combined into one set and stored in the *train_set* variable. Then, the remaining 20% of the data is used as testing data.

Then, the model validation process is carried out by training the Naive Bayes classification model using the training data. After that, the model's accuracy is calculated on the testing data by calling the 'accuracy ()' function and printing the results. This indicates the model's accuracy in classifying sentiment on unseen testing data. The following are the accuracy results of the Naive Bayes modeling for six marketplaces.

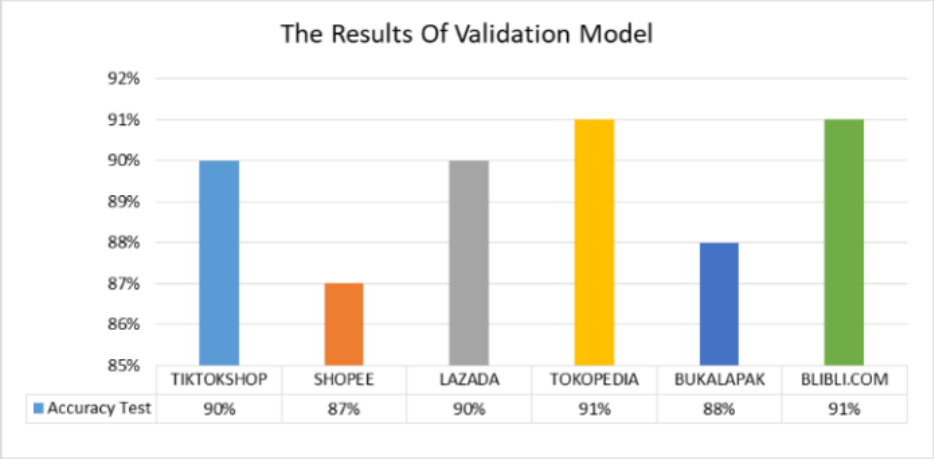
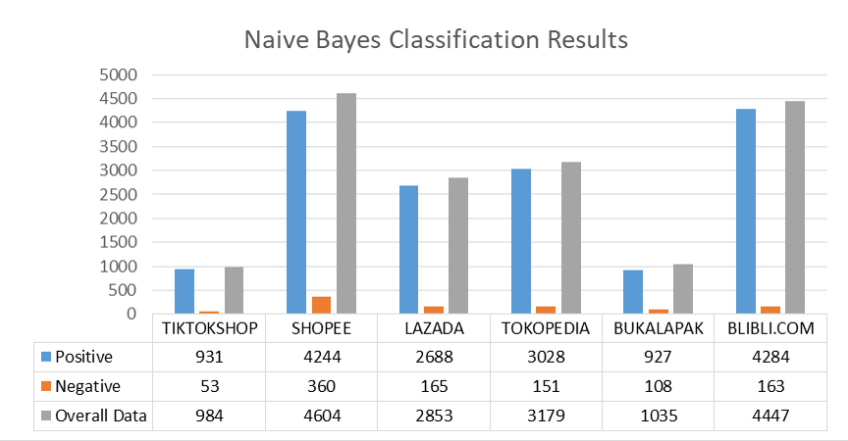


Figure 4 Model Validation Results

Based on the results of the validation model graph shown in Figure 4, it was found that the best accuracy was achieved by Blibli.com, Tokopedia, and Tiktokshop marketplaces with an accuracy score of 0.91 or 91%, followed by Lazada with an accuracy score of 0.90 or 90%. Bukalapak was in the fifth position with an accuracy score of 0.88 or 88%, and Shopee was in the last position with an accuracy score of 0.87 or 87%. From these results, it can be concluded that the validation of the naive Bayes classification method is reliable as the accuracy achieved is almost at the highest level.

In the next step, the tweet texts will be run through the Naive Bayes classification model to determine whether the tweet has a positive or negative sentiment. After that, the total number of tweets with positive and negative sentiments will be calculated from all data tweets. The following results from sentiment calculation using the Naive Bayes model for the 6 Marketplaces.



	TIKTOKSHOP	SHOPEE	LAZADA	TOKOPEDIA	BUKALAPAK	BLIBLI.COM
Positive	931	4244	2688	3028	927	4284
Negative	53	360	165	151	108	163
Overall Data	984	4604	2853	3179	1035	4447

Figure 5 Results of Naive Bayes Classification

Figure 8 above shows the graphical output of naive Bayes classification methods. The next step involves adding a new column containing the results of the Naive Bayes classification to the dataset, which aims to facilitate the analysis of the overall classification results. The following are the results of the Naive Bayes classification.

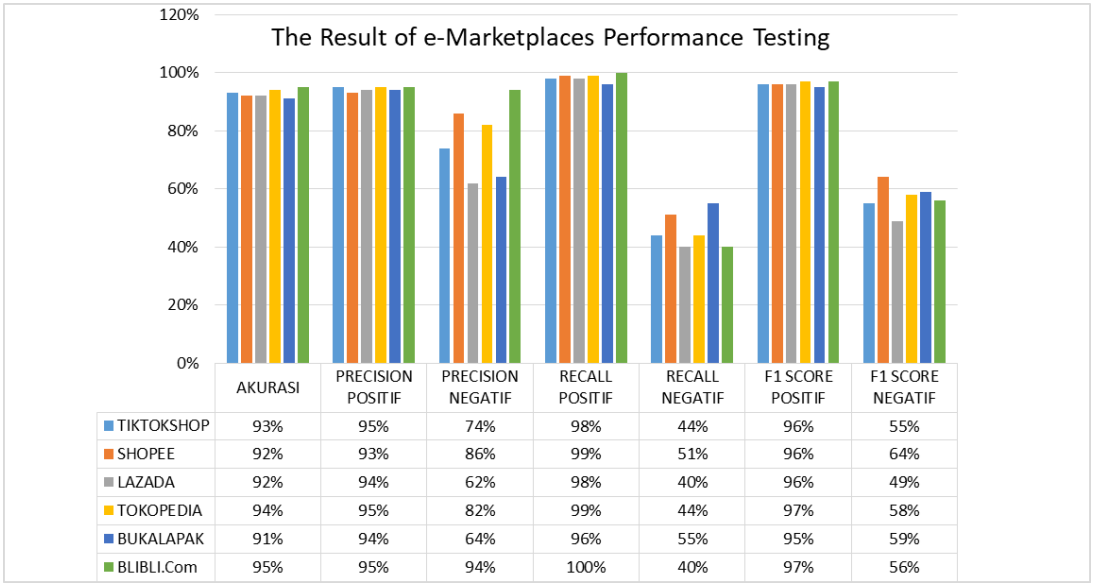
Table 11 Naive Bayes Classification Results

Sentiments	Klasifikasi_bayes
Positive	Positive
Negative	Negative
Negative	Positive

From Table 11, it can be seen that there is a difference between the labeling results using lexicon-based and Naive Bayes models on the third dataset in the table. Therefore, evaluation testing is needed to determine the accuracy and reliability of the prediction results generated by a model.

6. Metrics Performance Test Result

Figure 6 shows that the average accuracy of the six marketplaces is 0.93 or 93%, with the highest accuracy obtained by Blibli.com with a value of 0.95 or 95%, followed by Lazada with an accuracy of 0.94 or 94%.



	AKURASI	PRECISION POSITIF	PRECISION NEGATIF	RECALL POSITIF	RECALL NEGATIF	F1 SCORE POSITIF	F1 SCORE NEGATIF
TIKTOKSHOP	93%	95%	74%	98%	44%	96%	55%
SHOPEE	92%	93%	86%	99%	51%	96%	64%
LAZADA	92%	94%	62%	98%	40%	96%	49%
TOKOPEDIA	94%	95%	82%	99%	44%	97%	58%
BUKALAPAK	91%	94%	64%	96%	55%	95%	59%
BLIBLI.Com	95%	95%	94%	100%	40%	97%	56%

Figure 6 Performance Test Results

In the last position, Bukalapak obtained an accuracy value of 0.91 or 91%. The precision results of the model have an average of 0.94 or 94%, with the highest precision value obtained by the Blibli.com, Tokopedia, and Tiktokshop marketplaces with a result of 0.95 or 95%, and the lowest precision value obtained by Shopee with a result of 0.93 or 93%.

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 19, No. 2 September 2023 | DOI: 10.33480/pilar.v19i2.4176

The result of Recall from the model has an average value of 0.98 or 98%, with the highest Recall result obtained by Blibli.com with a value of 100% and the lowest value obtained by Bukalapak with a result of 0.96 or 96%.

The F1 Score results of the model have an average value of 0.96 or 96%, with the highest F1 Score obtained by Blibli.com and Tokopedia with a score of 0.97 or 97%, while Bukalapak obtains the lowest value with a score of 0.95 or 95%.

7. Visualization Test Result

Based on the visualization results in Figure 7, it can be concluded that the results of positive and negative sentiments from the six marketplaces show that Blibli.com obtained the highest positive sentiment with a result of 96.33%, and Bukalapak obtained the lowest positive sentiment with a result of 89.57%.

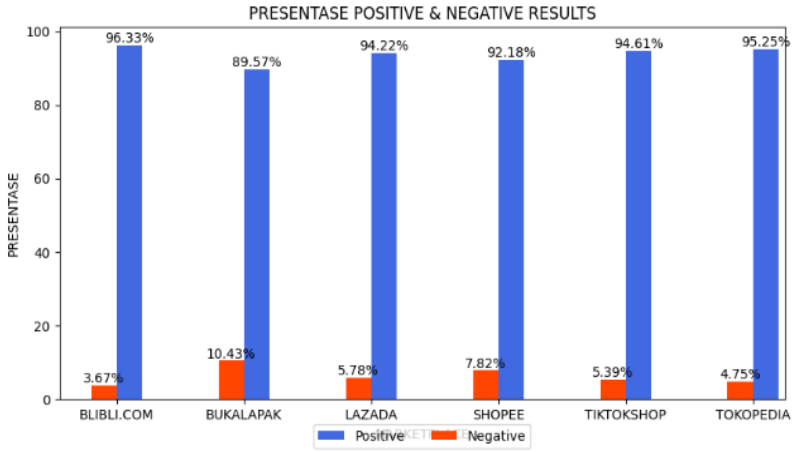


Figure 7 Presentase Positive & Negative Sentiments

Furthermore, the highest negative sentiment value is obtained by Bukalapak with a result of 10.43%, and Blibli.com obtains the lowest negative sentiment with a result of 3.67%.

Figure 8 Positive Review Visualization Results

Figure 8 shows that the most frequent positive words in the tweet data are *'Tiktokshop'*, *'Tiktok'*, *'real Indonesian'*, *'good'*, *'low price'*, and *'easy use'*.

CONCLUSION

The sentiment analysis related to selecting the best marketplace as a marketing strategy for MSMEs has been studied using customer reviews on Twitter for six marketplaces: Shopee, Tiktokshop, Lazada, Tokopedia, Blibli.com, and Bukalapak. The analysis involved pre-processing and weighting stages, labeling using the lexicon-based method, and classification with Naive Bayes machine learning. The results showed that the best accuracy was achieved by Blibli.com with an accuracy rate of 95% or 0.95, followed by Tokopedia with an accuracy rate of 94% or 0.94, Tiktokshop with an accuracy rate of 93% or 0.93, and Shopee and Lazada with an accuracy rate of 92% or 0.92. Bukalapak ranked last with an accuracy rate of 91% or 0.91.

In addition, the highest percentage of positive reviews was obtained by the blibli.com marketplace, with a positive sentiment of 96.33%. In second place was Tokopedia with a positive sentiment of 95.25%, followed by tiktokshop with a sentiment of 94.61%, then Lazada with a result of 94.22%, followed by Shopee with a result of 92.18%, and lastly Bukalapak with a result of 89.57%. Therefore, it can be concluded that in this research, the best marketplace for MSMEs marketing strategy, which is highly recommended, is blibli.com, followed by Tokopedia and tiktokshop. However, this study used different amounts of data for each marketplace, and the varying amounts of data can affect the sentiment analysis results. The more data obtained from a marketplace, the more accurate the sentiment analysis results will be.

After obtaining these results, it can be seen that there are differences between this study and previous studies. The use of a combination of the

two methods can improve the results in accuracy, precision, and recall, as well as the F1 score.

Conversely, the fewer data obtained, the less accurate the sentiment analysis results will be. Therefore, obtaining balanced data from each marketplace is essential to produce more accurate and reliable sentiment analysis results. Furthermore, to improve the quality of the results of this research, it is recommended to make advancements such as adding new, more specific, and relevant features that can be used for sentiment classification, conducting feature extraction using more sophisticated methods, or using other machine learning techniques. Additionally, manual labeling can also be employed, and the results of both methods can be compared to improve the performance of the methods and optimize the results to match the natural context.

REFERENCE

Ahdiat, A. (2022). Indonesia Punya UMKM Terbanyak di ASEAN, Bagaimana Daya Saingnya? Retrieved October 11, 2022, from databooks.katadata.co.id website: https://databooks.katadata.co.id/datapublish/2022/10/11/indonesia-punya-umkm-terbanyak-di-asean-bagaimana-daya-saingnya

Amaliah, F., & Nuryana, I. K. D. (2022). Perbandingan Akurasi Metode Lexicon Based Dan Naive Bayes Classifier Pada Analisis Sentimen Pendapat Masyarakat Terhadap Aplikasi Investasi Pada Media Twitter. *JINACS (Journal of Informatics and Computer Science)*, *3*(3), 384–393.

Asri, Y., Suliyanti, W. N., Kuswardani, D., & Fajri, M. (2022). Pelabelan Otomatis Lexicon Vader dan Klasifikasi Naive Bayes dalam menganalisis sentimen data ulasan PLN Mobile. *PETIR: Jurnal Pengkajian Dan Penerapan Teknik Informatika*, *15*(2), 264–275. https://doi.org/10.33322/petir.v15i2.1733

Darmawan, T. D. (2022). *Analisis Sentimen Areview Pelanggan E-Commerce di Indonesia Menggunakan Algoritma Naive Bayes*. Universitas Dinamika.

Doni. (2021). Pimpin Rapat Hilirisasi Ekonomi Digital, Presiden Instruksikan Percepatan Digitalisasi UMKM. Retrieved February 10, 2021, from kominfo.go.id website: https://www.kominfo.go.id/content/detail/34994/pimpin-rapat-hilirisasi-ekonomi-digital-presiden-instruksikan-percepatan-digitalisasi-umkm/0/berita

Faesal, A., Muslim, A., & Ruger, A. H. (2020). Sentimen Analisis pada Data Tweet Pengguna Twitter Terhadap Produk Penjualan Toko Online Menggunakan Metode K-Means. *Jurnal MATRIK*, *19*(2), 207–213. https://doi.org/10.30812/matrik.v19i2.640

Haranto, F. F., & Sari, B. W. (2019). Implementasi Support Vector Machine Untuk Analisis Sentimen Pengguna Twitter Terhadap Pelayanan Telkom dan Biznet. *Jurnal Pilar Nusa Mandiri*, *15*(2), 171–176. https://doi.org/10.33480/pilar.v15i2.699

Hartatik, Tamam, M. B., & Setyanto, A. (2020). Prediction for Diagnosing Liver Disease in Patients using KNN an Naive Bayes Algorithm. *International Conference on Cybernetics and Intelligent System (ICORIS)*. https://doi.org/10.1109/ICORIS50180.2020.9320797

Hasugian, A. H., Fakhriza, M., & Zukhoiriyah, D. (2023). Analisis Sentimen Pada Review Pengguna E-Commerce Menggunakan Algoritma Naïve Bayes Jurnal Teknologi Sistem Informasi dan Sistem Komputer TGD. *Jurnal Teknologi Sistem Informasi Dan Sistem Komputer TGD*, *6*(1), 98–107. Retrieved from https://ojs.trigunadharma.ac.id/index.php/jsk/index%0AAnalisis

Johnson, M., & Smith, K. (2022). Exploring Social Media Data Analysis Techniques: A Review of Netlytic. *Journal of Social Media Research*, *10*(1), 20–35.

Kemp, S. (2021). Digital 2021 : Indonesia. Retrieved February 18, 2023, from Datareportal We Are House website: https://datareportal.com/reports/digital-2021-indonesia

Kurniawan, A., Adinugroho, S., & Features, B. (2019). *Analisis Sentimen Opini Film Menggunakan Metode Naïve Bayes dan Lexicon Based Features*. *3*(9), 8335–8342.

Mahdi, M. I. (2022). Berapa Jumlah UMKM di Indonesia? Retrieved January 19, 2022, from dataindonesia.id website: https://dataindonesia.id/sektor-riil/detail/berapa-jumlah-umkm-di-indonesia

Mardiana, T., Syahreva, H., & Tuslaela. (2019). Komparasi Metode Klasifikasi Pada Analisis Sentimen Usaha Waralaba Berdasarkan Data Twitter. *Jurnal Pilar Nusa Mandiri*, *15*(2), 267–274. https://doi.org/10.33480/pilar.v15i2.752

Pamungkas, R. B. (2023). Hati-Hati Inilah Tantangan UMKM di Indonesia Pada Tahun 2023. Retrieved January 5, 2023, from Niagahoster.co.id website: https://www.niagahoster.co.id/blo/tantangan-umkm-indonesia/

Pratiwi, S. Y. A., & Nudin, S. R. (2021). Analisis Sentimen terhadap Facebook Marketplace

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 19, No. 2 September 2023 | DOI: 10.33480/pilar.v19i2.4176

Menggunakan Metode Lexicon Based dan Support Vector Machine. *JIFTI-Jurnal Ilmiah Teknologi Informasi Dan Robotika*, *3*(2), 9–15.

Rianti, D. L., Umaidah, Y., & Voutama, A. (2021). Tren Marketplace Berdasarkan Klasifikasi Ulasan Pelanggan Menggunakan Perbandingan Kernel Support Vector Machine. *STRING (Satuan Tulisan Riset Dan Inovasi Teknologi*, *6*(1), 98–105.

Risnasari, M. (2022). *Konsep Dasar Data Mining Teori dan Praktik dengan PYTHON* (1st ed.). Malang: CV Literasi Nusantara Abadi.

Tamam, M. B., Hozairi, Walid, M., & Bernado, J. F. A. (2023). Classification of Siggn Language in Real Time Using Convolutional Neural Network. *Applied Information System and Management (AISm*, *6*(1), 39–46.

Umbu, A., Ama, T., Mulya, D. N., Astuti, Y. P. D., Bias, I., & Prasadhya, G. (2022). Analisis Sentimen Customer Feedback Tokopedia Menggunakan Algoritma Naïve Bayes. *Jurnal Sistem Komputer Dan Informatika (JSON)*, *4*(September), 50–55. https://doi.org/10.30865/json.v4i1.4783

Utama, H. S., Rosiyadi, D., Aridarma, D., & Prakoso, B. S. (2019). Sentimen Analisis Kebijakan Ganjil Genap di Tol Bekasi Menggunakan Algoritma Naive Bayes dengan Optimalisasi Information Gain. *Jurnal Pilar Nusa Mandiri*, *15*(2), 247–254. https://doi.org/10.33480/pilar.v15i2.705

Wilandini, D., & Purwantoro. (2022). Penerapan Algoritma Naive Bayes dalam Mengklasifikasikan Media Sosial Untuk Mengamati Trend Kuliner. *Jurnal Teknologi Terpadu*, *8*(1), 31–39. Retrieved from https://journal.nurulfikri.ac.id/index.php/jtt

Wirma, S. (2022). Data Mining Dengan Metode Naïves Bayes Classifer dalam Memprediksi Tingkat Kepuasan Pelayanan Dokumen Kependudukan. *Jurnal Informatika Ekonomi Bisnis*, *4*(3), 119–123. https://doi.org/10.37034/infeb.v4i3.155