Journal of Student Research Exploration

https://shmpublisher.com/index.php/josre

p-ISSN 2964-1691 | e-ISSN 2964-8246

Implementation of Lexicon-Based and SVM Methods in Sentiment Analysis of Sayurbox App Users

Raihan Muhammad Rizki Rahman[1], Budi Prasetiyo[2]

[1,2] Department of Computer Science, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received July 9, 2024
Revised April 2, 2026
Accepted April 13, 2026

Keywords:

Sayurbox app
Lexicon-based analysis
User experience
User sentiments
Support vector machine algorithm (SVM)

ABSTRACT

The ever-growing technology certainly produces a large amount of data, which can provide useful information if analyzed and used properly. The purpose of this research is to analyze user sentiment towards the Sayurbox application on the Google Play Store with a Lexicon-Based approach and the Support Vector Machine (SVM) algorithm. User review data is obtained through web scraping with a total of 16,468 reviews. After preprocessing and sentiment labeling, training and test data were divided. The results showed that SVM achieved accuracy, recall, and precision of 94%, 96%, and 96% respectively, with 9 prediction errors. The model tends to predict reviews as positive sentiment, indicating user satisfaction with Sayurbox's product service, delivery, quality, and price. The findings make a contribution to the understanding of user sentiment in e-commerce services and can assist Sayurbox in improving their user experience.

1. Introduction

Technology's rapid growth can assist society in a variety of ways. The ever-changing technology undoubtedly generates a significant amount of data, which might be important information if studied and used correctly. Because of the high number of people that use the internet nowadays, the amount of data generated is enormous (Big Data) [1]. Big data technology facilitates the processing of massive and complicated data sets. [2], [3], such that the processed data can yield useful results.

[1] ***Corresponding Author:***
Raihan Muhammad Rizki Rahman,
Department of Computer Science,
Universitas Negeri Semarang,
Sekaran, Gunungpati, Semarang, Indonesia.
Email: raihanmuhammad22@students.unnes.ac.id
DOI: https://doi.org/10.52465/josre.v4i1.391

Big data is currently being used extensively in mobile applications and websites. The Big Data principle, which involves managing a large amount of data in a short period of time, is required to run programs efficiently [3]. To make the program more useful to a large number of individuals, an intermediate that makes it easier to access is required. Google's Play Store [4] provides a variety of digital content categories, including games, applications, programs, music, and books.

The Play Store feature employed is a rating or review, which allows users to express their thoughts on the program that has been used [4]. Sayurbox is one of the offered applications. Sayurbox is a mobile application that allows you to buy and sell kitchen items online. Sayurbox simply delivers kitchen requirements, such as fresh vegetables and fruits from farmers, quality protein, cooking packages, and cut veggies, as well as a variety of other basic needs, by forming partnerships [5] with farmers.

To continue to adjust the application to the demands of users, additional study of user evaluations or opinions, such as sentiment analysis, is required [6]. Machine learning, Lexicon-Based, hybrid, and other techniques such as multimodal sentiment analysis, aspect-based approach, and transfer learning are examples of commonly used sentiment analysis models [7]. This study will classify sentiment using a Lexicon-based method and the Support Vector Machine (SVM) algorithm on Sayurbox application user review data from the Google Play Store.

The SVM approach was chosen since it is a well-known algorithm in classification applications. SVM also has a high generalization rate even with limited training data and can be applied to high-dimensional data [8]. The Lexicon-Based technique was chosen because it is simple to develop, does not require a large amount of training data, is quick to run, transparent, can be tailored to the application environment, lowers the danger of overfitting, and successfully manages imbalance in review data [9]. The goal of this research is to examine the performance of the Support Vector Machine (SVM) algorithm in sentiment categorization

2. Method

An experimental approach and quantitative methods were used in this research, to examine user perception regarding the Sayurbox application. The user review data was collected from the Google Play Store and then processed through various steps like - Data Cleansing, tokenization, stopword removal & Text Normalisation. To analyse this information, we resorted to two classification methodologies: Lexicon-Based - which uses a sentiment-based lexicon for scoring words of the written review; and SVM (Support Vector Machine) algorithm. The research workflow is illustrated in Figure 1, which outlines the sequentioal stages of data collection, preprocessing, transformation, data mining, and evaluation.

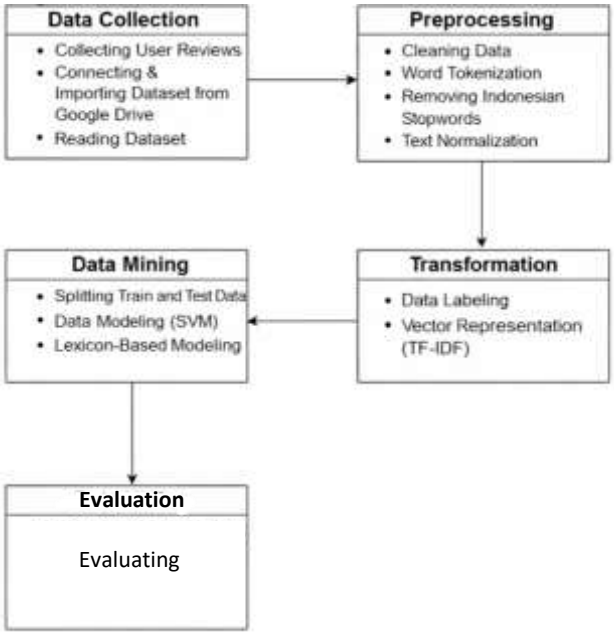

Figure 1. Flowchart of research method

2.1 Data Collection

The very first step in this research is data collection. This information was collected based on user reviews of the Sayurbox app in Google Play Store. The data collection process is through google colab using python and google play scraper module. Then connect and upload data sets from Google Drive so that you can add further information for a study. Once the dataset is imported, it also needs to be read and prepped up for computation.

2.2 Text Preprocessing

It then goes to the preperation phase, posing with data cleaning where all punctuation marks, digits and special characters not important for review are removed from it. Data preprocessing is a method to convert raw data in readable forms [10]. This is followed by word tokenization which separates the words in review content for further analysis. For stopwords removal, it is also used for Bahasa Indonesia to exclude frequent words that do not returns any significant meaning in sentiment analysis such as minimalistic bahasa-stopwords. In the end text normalization is used to transform terms into their root form making it easier for an analysis.

2.3 Transformation

Data Labeling: Assign sentiment labels (positive, negative or neutral) for each review based on a initial analysis, other data available in the form of you can ask someone to do so. Vector representatives (TF-IDF) is then used to convert the text reviews into a feature vector which allows us to build machine learning models.

2.4 Data Mining

The data mining approach first divides the training and test into a particular fraction to duly train & investigate the model. To achieve autonomous sentiment classification, data modeling utilizing the Support Vector Machine (SVM) methodology is applied for model training. It also adopts various models like Lexicon-Based Modelling which uses a lexicons to measure emotion score per review.

2.5 Evaluation

At the end of all this we reach to evaluation stage where performance of training model is judged. These metrics are used to measure the performance of model in terms of how well it can classify the positive and negative sentiments from reviews, confusion matrix, precision, recall, F1 score etc. Finally, we compare the sentiment analysis results using Lexicon-Based and SVM approach to find out which method gives higher accuracy for sentiment of Sayurbox application users.

3. Results and Discussion

userName	userImage	content	score	mbsUpCo	eviewCreatedVersior	at
Yuni Yuni	https://pl	pengiriman cepat..packing aman..barang fresh..ter	5	0	1.57.1	2022-03-31 23:22:32
Happy Agus Artawan	https://pl	sesuai pesanan dan kurirnya ramah mohon diperta	5	0	1.57.1	2022-03-31 23:18:27
Jakaria Amsari	https://pl	pengiriman tepat waktu , pilihan tepat berbelanja s	5	0	1.57.2	2022-03-31 22:54:45
novri apriyanto	https://pl	Cocok buat emak² mager cem saiaah 😅 udh gitu s	5	0	1.57.1	2022-03-31 08:01:26
Meni Chandra	https://pl	sayur box makin mantap aja nih ..pertahanan ya	5	0	1.57.1	2022-03-31 07:18:30
nida hanifah	https://pl	Thanks ya, promonya banyak. Saran aja mungkin b	5	0	1.57.1	2022-03-31 06:15:55

Figure 2. Table of Research Datasets

This study uses data from user reviews of the Sayurbox app on the Playstore. Data is collected using scraping techniques in Python on Google Colab, with supporting models such as nltk for scraping data from Google Play Store. The data that was successfully retrieved were user reviews from 2020 to 2022. When accessing Sayurbox application reviews, library reviews are used to obtain all of the information needed for analysis, such as reviewID, username, content, score, date, and so on as depicted in Figure 2.

Scraping techniques yielded a total of 1906 reviews or comments from Sayurbox application users. Following that, data duplication is cleaned, and the data is ready to move on to the next step, which is to clean the reviews of emojis, punctuation marks, font irregularities, word repetition, normalization, and so on as illustrated in Figure 4. The distribution of Sayurbox user ratings per year is illustrated in Figure 3.

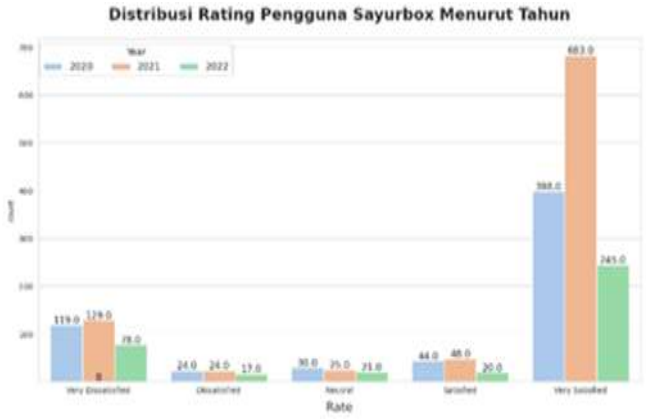

Figure 3. Sayurbox App user rating distribution by year

content	clean_review	normalization	final_text	token	stop_review	stem_review	
0	pengiriman cepat..packing aman..barang fresh.....	pengiriman cepat packing aman barang fresh ter...	pengiriman cepat pengepakan aman barang segar ...	pengiriman cepat pengepakan aman barang segar ...	[pengiriman, cepat, pengepakan, aman, barang, ...	[pengiriman, cepat, pengepakan, aman, barang, ...	kirim cepat epak aman barang segar terima kasi...
1	sesuai pesanan dan kurirnya ramah mohon dipert...	sesuai pesanan dan kurirnya ramah mohon dipert...	sesuai pesanan dan kurirnya ramah mohon dipert...	sesuai pesanan kurirnya ramah mohon dipertaha...	[sesuai, pesanan, kurirnya, ramah, mohon, dipe...	[sesuai, pesanan, kurirnya, ramah, mohon, dipe...	sesuai pesan kurir ramah mohon tahan
2	pengiriman tepat waktu , pilihan tepat berbela...	pengiriman tepat waktu pilihan tepat berbelanj...	pengiriman tepat waktu pilihan tepat berbelanj...	pengiriman tepat waktu pilihan tepat berbelanj...	[pengiriman, tepat, waktu, pilihan, tepat, ber...	[pengiriman, pilihan, berbelanja, sayuran, bua...	kirim pilih belanja sayur buah silah ragu bela...
3	Cocok buat emak² mager cem salaah 😭 udh gitu s...	cocok buat emak mager cemsalaah udh gitu sayu...	cocok buat ibu malas gerak macam saya sudah be...	cocok buat malas gerak macam saya sudah begit...	[cocok, buat, malas, gerak, macam, saya, sudah...	[cocok, malas, gerak, sayurnya, segar, murah, ...	cocok malas gerak sayur segar murah supermarke...
4	sayur box makin mantap aja nih ..pertahanan ya	sayur box makin mantap aja nih pertahanan ya	sayuran box makin mantap saja nih pertahankan ya	sayuran makin mantap saja pertahankan	[sayuran, makin, mantap, saja, pertahankan]	[sayuran, mantap, pertahankan]	sayur mantap tahan
...	...	...	...	...	...	...	...
617	Untuk beli beras apakah gratis ongkir??	untuk beli beras apakah gratis ongkir	untuk beli beras apakah gratis ongkos kirim	untuk beli beras apakah gratis ongkos kirim	[untuk, beli, beras, apakah, gratis, ongkos, k...	[beli, beras, gratis, ongkos, kirim]	beli beras gratis ongkos kirim
618	klannya mangga 1.000 per kg. Pas download sto...	iklannya mangga per kg pas download stok abis ...	klannya mangga per kg pas unduh stok habis ba...	klannya mangga unduh stok habis baik trik ...	[klannya, mangga, unduh, stok, habis, baik, t...	[klannya, mangga, unduh, stok, habis, trik, b...	iklan mangga unduh stok habis trik biar orang ...
619	Terima kasih sayur box, memudahkan saya belanj...	terima kasih sayur box memudahkan saya belanja...	terima kasih sayuran box memudahkan saya belan...	terima kasih sayuran memudahkan saya belanja ...	[terima, kasih, sayuran, memudahkan, saya, bel...	[terima, kasih, sayuran, memudahkan, belanja, ...	terima kasih sayur mudah belanja respons kompl...
620	Belanja jadi hemat	belanja jadi hemat	belanja jadi hemat mantap	belanja jadi hemat mantap	[belanja, jadi, hemat]	[belanja, hemat]	belanja hemat
621	Mntap smoga mnjadi lebih baik lgi untuk mmajuk...	mntap smoga mnjadi lebih baik lgi untuk mmajuk...	semoga menjadi lebih baik lagi untuk me...	semoga menjadi lebih baik lagi untuk me...	[mantap, semoga, menjadi, lebih, baik, lagi, u...	[mantap, semoga, memajukan, petani]	mantap moga maju tani

Figure 4. Data cleaning and tokenization

Data exchange occurs when text preparation and sentiment tagging are completed. Training data is used to train the algorithm, whereas testing data is used to evaluate the performance of the trained technique when new data is acquired that has not previously been gathered.

Before entering the data labelling process, comments lack sentiment, making it difficult to determine whether users provide favorable or negative feedback. Because sentiment tagging is difficult to conduct manually, particularly with huge datasets, a Lexicon-based method is utilized [11]. The application makes use of the InSet Lexicon dictionary [11], which includes both positive and negative dictionaries.

The reviews are then allocated a value that affects the weight of the terms in the dictionary. If it contains positive terms, the value increases by 5, while negative words reduce the value by 5. After that, the word weights are tallied and classified as positive, negative, or neutral. Neutral here indicates that the weight gained is equal to zero. The amount of positive and negative labels is illustrated in Figure 5, showing that positive labels dominate the dataset compared to negatives ones.

Amount of positive and negative labels

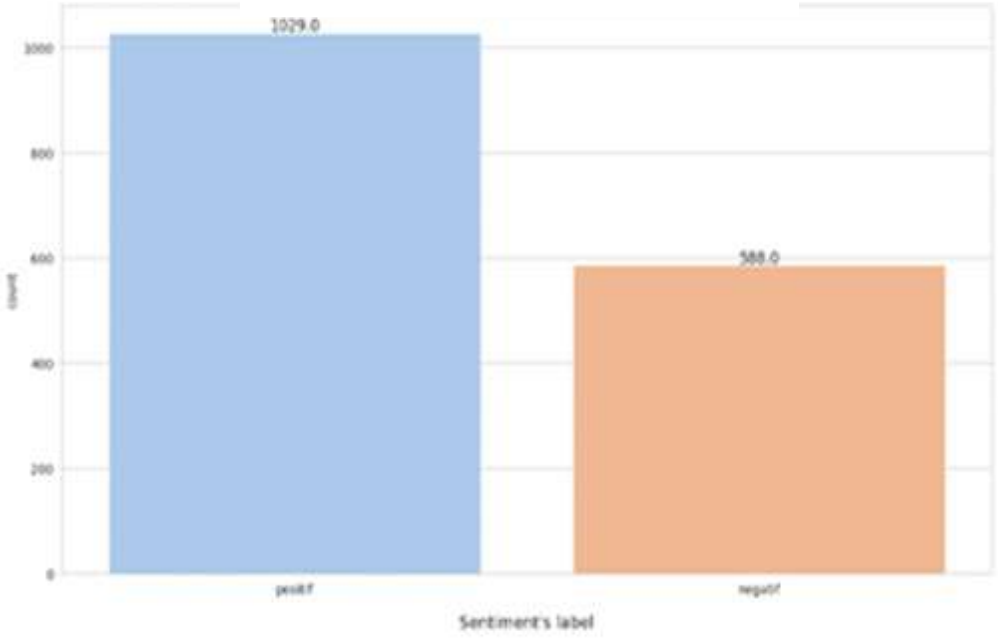

Figure 5. Amount of positive and negative labels

After labeling the data, the reviews are divided into training and testing sets. We also find the parameters or limitations for data splitting that produce the maximum accuracy values in order to reduce the incidence of model prediction mistakes.

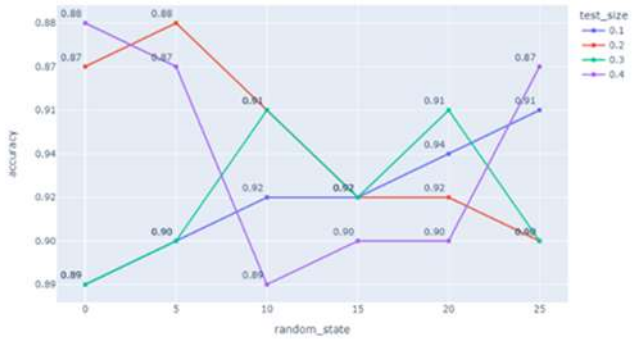

Figure 6. Visualization of accuracy graph

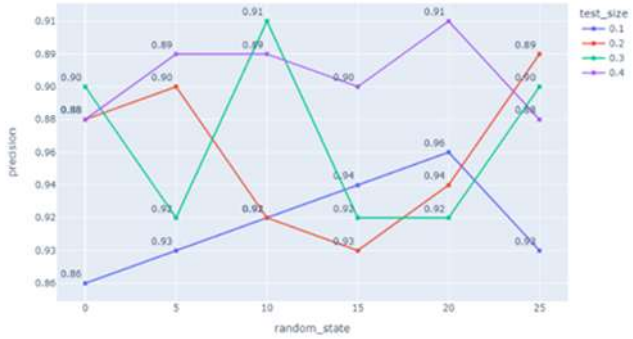

Figure 7. Visualization of precision graph

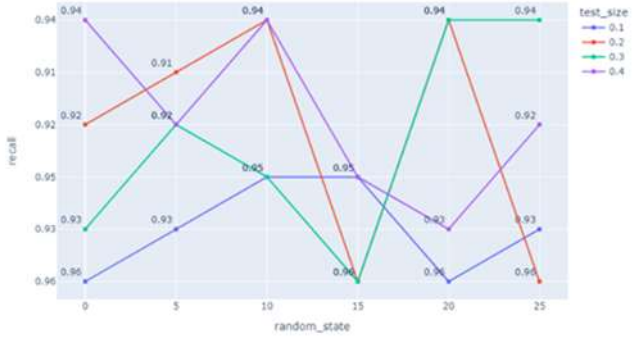

Figure 8. Visualization of recall graph

According to the findings of the preceding study into parameters or constraints on the optimal test_size and random_state, the weighting of test_size = 0.1 and random_state = 20 yields superior accuracy, recall, and precision values than other sizes. So the obtained size will be utilized to split the dataset. The visualization of precicion and recall graphs is presented in Figure 7 dan Figure 8, respectively, complementing the accuracy graph in Figure 6.

As a consequence, the model can predict positive sentiment review test data by 70.37% (114 reviews) and negative sentiment by 29.63% (48 reviews). It is clear that the value of positive sentiment is more widely distributed because the model predicts and defines more new reviews as positive sentiment than negative sentiment, and because the model previously learned and recognized more positive review data patterns than negative, it learned fewer negative review data patterns.

Sayurbox App Review sentiment analysis
prediction results

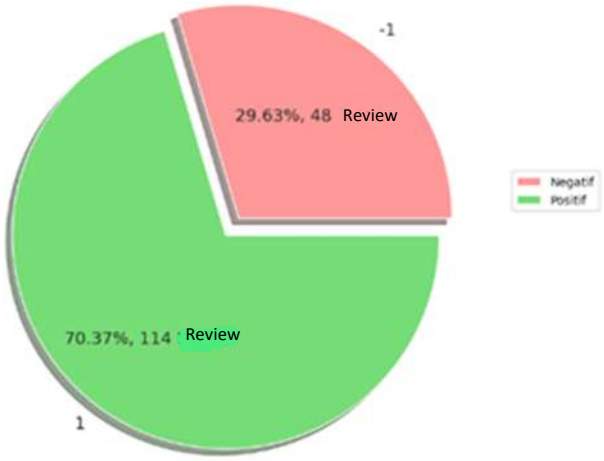

Figure 9. Sayurbox App Review sentiment analysis prediction results

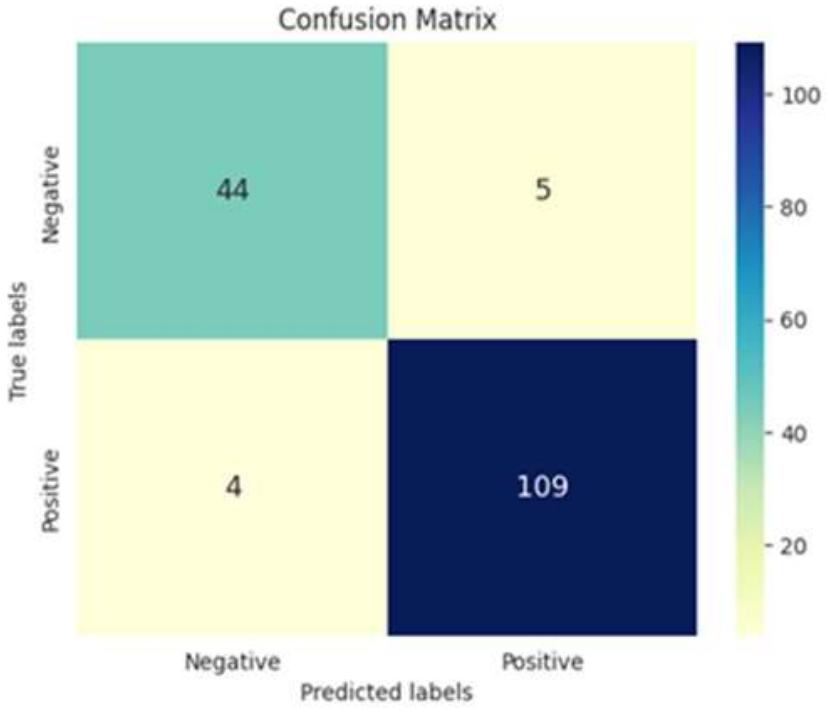

Figure 10. Confusion matrix

According to the research, the accuracy rate is 94%, recall is 96%, precision is 96%, and prediction errors are 9 data points. Furthermore, the Sayurbox review test data predicted 114 favorable reviews and 48 negative reviews. The Support Vector Machine Algorithm is recognized to forecast test data as positive sentiment rather than negative sentiment. As a result, it is reasonable to assume that Sayurbox's customers are satisfied with the service, delivery, quality, and price of its items.

4. Conclusion

According to the findings of this study, Sayurbox app users tend to provide good evaluations, indicating satisfaction with the service, delivery, product quality, and price. The sentiment analysis method, which employs the Lexicon-Based methodology and the SVM algorithm, produces satisfactory findings with high accuracy, recall, and precision. Although the algorithm predicts positive sentiment in evaluations, this can be used to help Sayurbox enhance the quality of their services and goods.

REFERENCES

[1] O. Solihin, "IMPLEMENTASI BIG DATA PADA SOSIAL MEDIA SEBAGAI STRATEGI KOMUNIKASI KRISIS PEMERINTAH," *Jurnal Common |*, vol. 5, no. 1, 2021, doi: 10.34010/common.

[2] V. Ferdiansyah, M. Irwan, and P. Nasution, "Penerapan Teknologi Big Data Dalam Pengembangan Database Pendidikan," *Jurnal Riset Manajemen*, vol. 1, no. 3, pp. 22–29, 2023, doi: 10.54066/jurma.v1i3.591.

[3] D. Kusumasari, O. Rafizan, and D. O. Rafizan, "STUDI IMPLEMENTASI SISTEM BIG DATA UNTUK MENDUKUNG KEBIJAKAN KOMUNIKASI DAN INFORMATIKA STUDI IMPLEMENTASI SISTEM BIG DATA UNTUK MENDUKUNG KEBIJAKAN KOMUNIKASI DAN INFORMATIKA Study on Implementation of Big Data System for Supporting Communication and Informatics Policy," *Jurnal Masyarakat Telematika dan Informasi* , vol. 8, pp. 81–96, 2017.

[4] N. C. Agustina, D. Herlina Citra, W. Purnama, C. Nisa, and A. Rozi Kurnia, "The Implementation of Naïve Bayes Algorithm for Sentiment Analysis of Shopee Reviews on Google Play Store Implementasi Algoritma Naive Bayes untuk Analisis Sentimen Ulasan Shopee pada Google Play Store," *MALCOM: Indonesian Journal of Machine Learning and Computer Science* , vol. 2, pp. 47–54, 2022.

[5] J. Ginanjar and I. Sukoco, "PENERAPAN DESIGN THINKING PADA SAYURBOX," *JURISMA: Jurnal Riset Bisnis dan Manajemen*, vol. 12, no. 1, 2022.

[6] G. Manik, I. Ernawati, and I. Nurlaili, "Analisis Sentimen Pada Review Pengguna E-Commerce Bidang Pangan Menggunakan Metode Support Vector Machine (Studi Kasus: Review Sayurbox dan Tanihub pada Google Play)," *Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA) Jakarta-Indonesia*, 2021.

[7] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowl Based Syst*, vol. 226, Aug. 2021, doi: 10.1016/j.knosys.2021.107134.

[8] T. Meisya, P. Aulia, N. Arifin, and R. Mayasari, "PERBANDINGAN KERNEL SUPPORT VECTOR MACHINE (SVM) DALAM PENERAPAN ANALISIS SENTIMEN VAKSINISASI COVID-19," *SINTECH Journal*, vol. 4, 2021, [Online]. Available: https://doi.org/10.31598

[9] R. Catelli, S. Pelosi, and M. Esposito, "Lexicon-Based vs. Bert-Based Sentiment Analysis: A Comparative Study in Italian," *Electronics (Switzerland)*, vol. 11, no. 3, Feb. 2022, doi: 10.3390/electronics11030374.

[10] D. ’ Rohannisa, F. Daud, B. Irawan, and A. Bahtiar, “PENERAPAN METODE NAIVE BAYES PADA ANALISIS SENTIMEN APLIKASI MCDONALDS DI GOOGLE PLAY STORE,” *Jurnal Mahasiswa Teknik Informatika*, vol. 8, no. 1, 2024.

[11] S. A. Ashari, M. W. A. Saputra, E. Larosa, and B. S. Rijal, “Analisis Sentimen pada Aplikasi Translate Google Menggunakan Metode SVM (Studi Kasus: Komentar Pada Playstore),” *Jurnal Teknik*, vol. 21, no. 2, pp. 168–182, Dec. 2023, doi: 10.37031/jt.v21i2.412.