Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13141
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Sentiment Analysis by Using Naïve Bayes Classification and Support Vector Machine, Study Case Sea Bank

Yefta Christian[1)], Tony Wibowo[2)], Mercy Lyawati[3)*]
[1,2,3)]Universitas Internasional Batam, Indonesia
[1)] yefta@uib.ac.id, [2)] tony.wibowo@uib.ac.id, [3)]2031147.mercylyawati@uib.edu

Submitted: Nov 2, 2023 | **Accepted** : Nov 6, 2023 | **Published** : Jan 1, 2024

Abstract: Information technology is developing at a rapid pace, changing people's lives, particularly in the financial sector where customer demands are rising, and banks must innovate to convert from traditional to technological banking systems while also increasing competency and efficiency through improved services. Innovations in digital banking have arisen in Indonesia as a result of technical progress. SEA Bank is one such digital bank; it was established in Indonesia in 2021. An app that may be found on the Google Play Store is used for all transactions. However, there are instances when the application's performance falls short of users' expectations, which prompts some users to voice their dissatisfaction. In order to determine if the evaluations are either beneficial or detrimental, the author therefore carried out a sentiment analysis study on SEA Bank using the Naïve Bayes classification and Support Vector Machine techniques. This was then implemented on a website utilizing the Flask framework. In the experiments with 90% training data, 10% testing data, and k = 10, the results of this study demonstrated that the sentiment classification process using the SVM algorithm was the best classification algorithm for evaluating its accuracy, precision, and recall values of 93.99%, 94.60%, 98.87%, and an F1 score of 96.69%.

Keywords : Flask Framework; Google play store; Machine Learning; SDLC; Sentiment analysis

INTRODUCTION

The development of information technology is increasing dramatically and made changes in people's lifestyles especially in financial industry. Banking efforts to generate profits cannot be separated from its ability to manage risk. As is known, risk as a fluctuation of results due to future uncertainty contains positive and negative dimensions. In the event of profit, the high fluctuation of results illustrates the opportunity for a large profit as well. However, in the event of a loss, then high fluctuations lead to large losses as well. Therefore, in the context of banking risk management, various efforts are made to minimize the negative effects of future uncertainty (Cahyaningrum & Atahau, 2021).

In Indonesia the Digital Bank innovations are emerging due to technological developments. Startups such as Tokopedia, Shopee, Gojek, Grab, and others have influenced the change in the world and have influenced the change in people's habits plus views. Where transactions from startups encourage the use of electronic transaction increasingly. Which caused the startup company with high technology to take advantages in the flaw in the service of conventional bank that are outdated and limitation in industry regulation, structure, and corporate culture (Cupian & Akbar, 2020).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i1.13141 p-ISSN : 2541-044X

The threat posed by financial start-ups or fintech is due to several things. Firstly, the growth of e-wallet or e- payment fintech has been quite rapid in recent years especially during the covid-19 such as Gopay, Shopee pay, and Ovo are the most popular electronic payments in Indonesia (Ramli, 2021). Second, the emergence of loan/financing innovations launched by various fintech. With innovation in the form of a fast and easy process, it will attract many people to try it. There are 125 fintech lenders that registered in OJK (*Otoritas Jasa Keuangan*) from 2017 to 2021. This factor is very influential because the main income of banks comes from the distribution of debt or financing.

Due to the above-mentioned factors, big corporate banks are transforming their current conventional banking system to digital banking where most transactions are used with an online system and saving under surveillance of OJK (*Otoritas Jasa Keuangan*).

According to Tony, the OJK Deputy Director of Basel and International Banking, several Indonesian banks have announced their full digital transformation, and more are in the process. Additionally, he disclosed that there will be 12 digital banks, five of which—Jago from Bank Jago, Wokee from Bank Bukopin, Digibank from DBS Bank, TMRW from Bank UOB, and Jenius from Bank BTPN—were founded as digital banks. The remaining seven banks—Bank BCA Digital, BRI Agroniaga, Bank Neo Commerce, Bank Capital, Bank Harda Internasional, Bank QNB Indonesia, and KEB Hana Bank—are in the process of converting to digital banking. According to Tony, there are distinctions between digital and pure banking. Digital banks involve BCA Digital, BRI Agro, Jenius BTPN, Digibank DBS, and others that have a relatively strong banking capital, particularly if they are owned by large banks. These have a more comprehensive digital ecosystem and services, such as Bank Jago, Line Bank, Sea Bank, and others (Ridhoi, 2021).

The digital banks that author would like to analysis is Sea Bank. SEA Bank, they are first built in 1991 as PT. Bank Kesejahteraan Ekonomi (Bank BKE) and officially changed the name to PT. Bank SeaBank Indonesia in 2021 (seabank.co.id, 2021). The main interest for citizen in choosing SEA Bank are due to the appealing high interest rates for savings and deposits in the range of 5% to 6%. Which make the digital banking more appealing than using the conventional banks. However, application performance sometimes is not as the user expectations, and this make some users have opinions and disappointed. Due to that some of the users think that the application is useless and hard to operate, although there are a lot of advantages. Thus, in Google Play store, its provided comments field for user reviews.

The google play store compile different feedback from users whether negative or positive feedback which make it difficult to determine if this application is good to use or not due to a complex system. With classification we can determine the review from play store if it's positive or negative and it will be much easier for developer for improving the application based on the user feedback, in addition there can be a comparison of application for users to choose what application suit them better for their need (Fransiska & Irham Gufroni, 2020).

Due to above mentioned risk and causes the author tempted to do a sentiment analysis, the foundation of machine learning and natural language processing, which is the field of text mining research that is currently trending. It is a crucial decision-making source that may be located, extracted, and assessed from online emotive reviews (Shamantha Rai B & Shetty Sweekriti M, 2019), specifically for SEA Bank by using google play scrap where author will make a python module for gathering data as for the analysis the author will use SVM and NBC algorithm for analyzing data from Google play store review of SEA Bank.

SVM is for finding the best hyper line to divide two types of classes and NBC is a straightforward technique that performs well and has high accuracy in text classification (Wahyu Handani, Intan Surya Saputra, Hasirun, Mega Arino, & Fiza Asyrofi Ramadhan, 2019). Both methods are being used to determine the accuracy of the final data information. Upon received the analyzed data of both method author will do an implementation by develop sentiment analysis website.

LITERATURE REVIEW

Research that was done by (Maulana, Susanto, & Kusumaningrum, 2019) is about developing a web scraping platform for the Indonesian market. The goal of this project is to help drop shippers obtain and

*name of corresponding author

automatically upload product data. The Laravel framework and waterfall development methodology are being used on this project. This study results in an application that can scrape products from a supplier's store, retrieve data, and then receive the collected data in a.csv file. Additionally, uploading is completed automatically and only requires the file name that was extracted from the.csv file to be entered. Retrieving product data from Tokopedia and Shopee marketplaces and uploading it to specific e-commerce platforms allows this project to be completed successfully.

Research of (Ilmawan & Mude, 2020), with the title of Comparison of Support Vector Machine and Naïve Bayes Classification Methods for Sentiment Analysis in Textual Reviews on Google Play Store. Using the NBC method, the effectiveness of the SVM classification method will be compared with other classification methods in this study. According to some earlier studies, the NBC method is a light classification technique with high accuracy when applied to text classification. The K-fold cross validation method is used to gauge the classifier's accuracy; the results are tabulated in a confusion matrix table, with a value of K = 3. When compared to the accuracy of the NBC, the SVM classifier achieves an accuracy of 81.46%, while the Naïve Bayes classifier receives an accuracy of 75.41%, according to the test results obtained from the 3-fold cross-validation method.

Research by (Nur'aini & Alfirman, 2021) regarding the sentiment analysis of WhatsApp's new privacy and policy regarding SVM and NBC. The objective is to ascertain the proportion of users who are in favor of WhatsApp's new privacy policy and to evaluate the effectiveness of the two algorithms. The training and testing data are divided into 90%:10%, 70%:30%, 50%:50%, 30%:70%, and 10%:90% groups in this study utilizing the NBC and SVM methods. The results showed that there were 338 positive reviews and 422 negative reviews, leading to the conclusion that, with a percentage of 56%, users were generally unhappy with WhatsApp's new privacy policy.

Research of (Salehudin Basryah, Erfina, & Warman, 2021) was analysis the sentiment of E-Money (Digital wallet) application in play store during the pandemic. The purpose of this research is to have an understanding whether the rating is reliable compared to the reality feedback. This study is using the NBC algorithm. Conclusion of this research is the highest accuracy value of the digital wallet application is Payfazz but has low positive review sentiment of compared to Dana. Therefore, Dana application is a recommended digital wallet as it has most positive sentiments.

Next in the research of (Aditra Pradnyana, Gede, & Darmawiguna, 2021) with the title of web-based system for Bali Tourism Sentiment Analysis. The goal of this study is to support the government's analysis of public sentiment regarding Bali tourism during the pandemic. The website was created using the Django framework and waterfall development method, while a web-based system was created to analyze the sentiment of Bali tourists using the Naïve Bayes method. The findings of this study indicate that, out of the total 2377 tweets that were collected between March 1st, 2020 and March 1st, 2021, there were still 65.67% tweets that were positive and 34.33% tweets that were negative during the pandemic.

In the research (Rudra Kumar, Pathak, & Gunjan, 2022) with its title of Diagnosis and Medicine prediction for Covid-19 using machine learning approach with classification model of CNN prediction with precision of 96% on the test set and 97% on the validation set and develop application using Flask framework.

Last and not least, research by(Alawi, Jamil Jastini Mohd, & Shaharanee, 2022), this paper aimed to predict student performance by analyzing the Oman student portal by employing J48 Decision tree algorithm with SEMMA methodology. This case study shows results this J48 Decision tree has prediction accuracy of 98.6%.

Table 1. Literature Review

Year	Author	Conclusion

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i1.13141 p-ISSN : 2541-044X

2019	Maulana, Susanto and Kusumaningrum	The research developed a web scraping in Indonesia Marketplace using the development method of waterfall and Laravel framework so the drop shipper can upload the catalogue automatically to their shop which is successfully done.
2020	Ilmawan and Mude	This research was done by doing a comparative review of 2 algorithm which is Naïve Bayes and SVM with the result show that SVM is more excel than NBC by using the accuracy of K-Fold validation of K=3.
2021	Nur'aini and Alfirman	The research was to analyze WhatsApp new privacy policy by using the NBC and SVM by splitting the data into 5 combinations of testing and training, and the model performance is tested with confusion matrix and accuracy as parameter and its shown that its best using SVM as its more accurate.
2021	Salehudin Basryah, Erfina and Warman	The research was about sentiment analysis of E-Money (Digital Wallet) by comparing 5 digital wallet application in play store and gathering information using web crawler and analysis with Naïve Bayes Method with accuracy as parameter. The result show that accuracy can't be the only parameter for sentiment analysis.
2021	Pradnyana and Darmawiguna	This study focused on creating a web-based platform that employs the waterfall method and Django framework to analyze public opinion in Bali tourism. The collected data is then subjected to naïve bayes analysis. The Naïve Bayes method is tested and evaluated using 5-Fold Cross Validation, which yields values for the confusion matrix, accuracy, precision, recall, and F-measure.
2022	M. Rudra Kumar, Vinit Kumar Gunjan and Rashmi Pathak	This research is about developed an application for covid-19 detection by using machine learning method with CNN prediction model and develop the application using the Flask Framework.
2022	Alawi, Jamil and Shaharanee	This research was aiming in analyzing Oman student performance by employing J48 Decision tree algorithm by using with SEMMA methodology and the model was validated using 10-fold cross with parameter confusion matrix, precision, recall, F-measure.

This research the author is focusing in analyze sentiment of SEA Bank by gathering the data using google play scraper module, The collected data from python module will be analyzed by using data mining process (framework) - SEMMA methodology like the research of (Alawi et al., 2022) by employing algorithm model of Naïve Bayes Classification and Support Vector machine like the research of (Nur'aini & Alfirman, 2021). Parameter of this research are confusion matrix, accuracy, precision, recall and F-measure as the past research of (Aditra Pradnyana et al., 2021). After the analysis , author will develop a web application for showing sentiment of each inputted text with waterfall development method (Maulana et al., 2019) and using the Flask Frame Work like the past research of (Rudra Kumar et al., 2022).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13141
e-ISSN : 2541-2019
p-ISSN : 2541-044X

METHOD

This project will be carried out in a systematic structure which is research flow. Research flow is used to explain the overall research from initial stage to finish stage that will be done by the author. Please refer to figure 1.

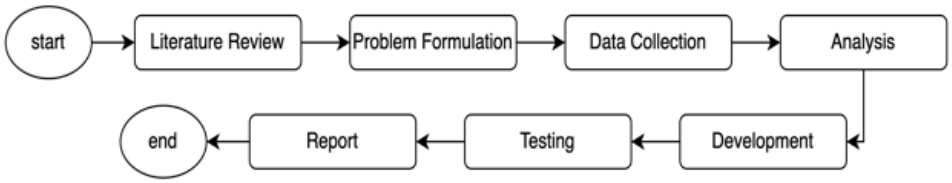

Fig. 1 Research Flow

This research is begun with gathering all related literature review and ground theory that related to the project topic which will become references for this research. During the problem identification phase, author will compile the problem indicator which will be solved with this research. After problem identification, author will collect data which shall be done by using the google play scrap, once the needed data had been gathered Author will do a data analysis by using data mining process (framework) – SEMMA with algorithm SVM and NBC algorithm for answering the indicated problems. In the last stage, author will write report of the research result by develop an analysis sentiment website based on the collected data with waterfall development method and website testing by using black box method.

Problem Formulation

This problem formulation is based on previous research that been carried out by (Alawi et al., 2022), (Nur'aini & Alfirman, 2021), (Aditra Pradnyana et al., 2021), (Maulana et al., 2019), and (Rudra Kumar et al., 2022), This research will be done by analyze collected data with data mining process (framework) – SEMMA methodology by using SVM and NBC algorithm and to have an understanding on algorithm accuracy in classifying Data with parameter of confusion matrix, accuracy, precision, recall and F-measure than developed a website with waterfall development method for showing SEA Bank sentiment analysis from play store.

Data Collection

The data collection of this research will be done by using the google play scraper. The information will be gathered from the play store web, based on the date, user rating and feedback of the SEA Bank. The targeted data of the SEA Bank will be from January 2021 to October 2023.

Analysis Data

Author will use SEMMA data mining framework to analyze the collected data. SEMMA is consisting of 5 phases please refer to figure 2 and further explanation.

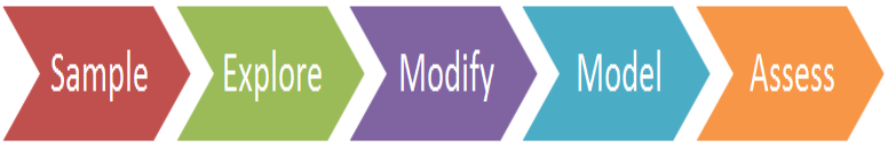

Fig. 2 Semma Methodology

a. Sample
The sample data of this staged are collected during the data collection phase that will be done from
January 2021 to October 2023, and the gathered data will be split into training and testing sets, with each set of data using five cycles of ratio for training and testing refer to table 2.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i1.13141 p-ISSN : 2541-044X

Table 2. Data Split

No.	Data Training	Data Testing	Total
1	90%	10%	100%
2	70%	30%	
3	50%	50%	
4	30%	70%	
5	10%	90%	

b. Explore

The next step is to explore the data visually or numerically for inherent grouping and helps in refine & redirect the discovery process to gain understanding.

c. Modify

This stage is focusing on model construction process. By using the patterns that are discovered during the exploration phase. In this stage the author will pre-process the collected data into 5 stages which are:

1. Case folding is the stage where all letter is changed to lowercase.
2. Data Cleansing is the process of removing noise such as emoticon or character that is not important in review sentences.
3. Tokenizing is the process of tokenizing or splitting a string, text into a list of tokens.
4. Lemmatization is the process convert words into its meaningful base form.
5. Stop word the process of eliminating words that fall under the category of stop words, words that frequently appear but have no meaning.

d. Model

Following the modification phase, the author will use the TF-IDF formula to vectorize the original textual data into numerical data and then use the SVM and NBC algorithms to construct the classification models. Please find the model equation mentioned below.

TF IDF Equation

The TF-IDF will be calculated with the following formulas:

$$TF\ (t,d) = \frac{number\ of\ times\ t\ appears\ in\ d}{total\ number\ of\ terms\ in\ d} \quad (1)$$

$$IDF\ (t) = log\frac{N}{1+df} \quad (2)$$

$$TF - IDF\ (t,d) = TF\ (t,d) * IDF\ (t) \quad (3)$$

Abbreviation:
d = document
N = total number of documents
df = number of documents with term t
TF-IDF = word frequency scores

NBC Equation

Naive Bayes is an algorithm that computes probabilities by considering conditional probabilities. Consequently, its formula relies on the characteristics of your analyzed data class and features. This approach yields a method known for its strong accuracy and performance in text classification.

$$P\ (C_k \mid X) = \frac{P\ (X \mid C_k) * P(C_k)}{P(X)} \quad (4)$$

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13141
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Abbreviation:
P = Probability
C = Class (Positive and Negative / label)
X = Features (Lemmatized review)

SVM Equation

The process of evaluation involves identifying decision boundaries or hyperplanes that divide a class from the others. In this instance, these boundaries help distinguish between sentences expressing positive (marked +1) and negative (marked -1). Thus, the SVM formula appears below.

$$(w.x_i) + b = 0 \quad (5)$$

When the information falls into class 1, it can be expressed as in the following.

$$(w.x_i) + b \leq 1, y_i = -1 \quad (6)$$

And the subsequent equation will apply to the data in class +1.

$$(w.x_i) + b \leq 1, y_i = 1 \quad (7)$$

e. Assess

Lastly, author will evaluation the performance results of the proposed classification SVM and NBC by calculating confusion matrix to get accuracy, precision, recall and F-measure refer to equitation 8. Once it's done, author will make conclusion of the model accuracy and comparison review of SVM & NBC based on the obtained results.

Confusion Matrix Formula

Confusion matrix for each classifier operation that utilizes True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN) components can be found using the following formula:

Table 3. Confusion Matrix Terminology

	Positive	**Negative**
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Accuracy Calculation

It's determined by dividing the total number of correct predictions by the number of datasets.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (8)$$

Precision Calculation

The precision metric indicates the proportion of accurately predicted cases that resulted in positive outcomes.

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13141
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Recall Calculation
Recall indicates the proportion of real positive cases that our model was able to accurately predict.

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

F1-Score Calculation
The F1-score provides a combined understanding of Precision and Recall as a mean of these two metrics. When Precision and Recall are equal, it is maximum.

$$F1 - Score = \frac{2}{\frac{1}{Recall}+\frac{1}{Precision}} \quad (11)$$

Development and Testing

Author will develop the website by using waterfall method. This method is being used since the direct purpose for this website is for show case result of based on the data that had been analyze. Waterfall development method consisting of 6 stages refer to figure 3 and define based on below statement:

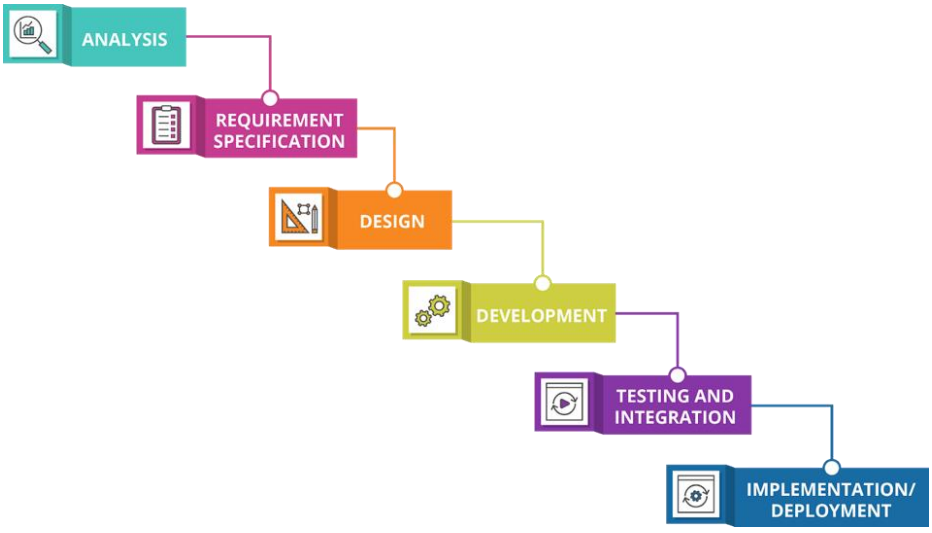

Fig. 3 Waterfall Methodology

a. Analysis & Requirement Specification.
In this phase the author will analyze and gather required information to develop the desired website. This system model will be explained with unified modeling language (UML). Use case will be used to understand the main feature or flow of the system and activity diagram will be presented to understand how the system overall works in structure.
b. Design
In this design phase the author will make sketch of the website user interface and database based on the system model which will become the reference during the development progress.
c. Development
During the development, the author will develop a website based on the system model and design that had been made previously, into the software with python programming language and Flask frameworks.
Below is the hardware specification that will be used by the author to develop the system:

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13141
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 4. Hardware Specs

Nama	Keterangan
Komputer	Macbook Air (13 inch, 2020)
Prosesor	1.1 GHz Quoad-Core Intel Core i5
Memori	8 GB 3733 MHz LPDDR4X
Kartu Grafis	Intel Iris Plus Graphics 1536 MB
MacOS	Ventura 13.3

The software requirement will be per below list:
1. Visual Studio Code (1.8.3.1)
2. Google Chrome (118.0.5993.117)
3. Jupyter Notebook (6.5.4)

Testing and Integration

In the testing phase, author will do the trial by using black box testing method. The black box testing method is being used to test the software in functionality. Author will input comment and check whether it can show the required analyze data based on variable. The testing is being done to validate whether the function of the input and output of the web scrapping tool is corresponding to the programmed structure.

Implementation/Deployment

The implementation/deployment of the website will be based on the analyzed data. This includes launching the website that is ready for use after going through the development and testing stages.

Report

The outcome of this project will be a sentiment analysis website based on the analyzed data and articles for documentation. This report will encompass the analysis results, a description of the website, and documentation regarding the methods and tools used in the research. This will aid in understanding and communicating the research findings to interested readers.

RESULT

Data Collection

The data is collected by using google play scraper by using a python module which is shown in figure.4.1.

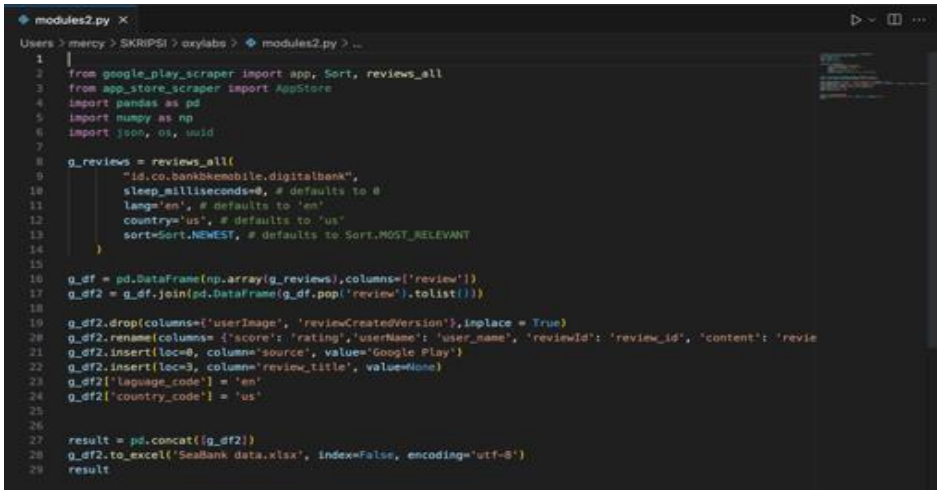

Fig. 4 Python Module

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13141
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Author is using google play scraper for getting review of SeaBank application with the parameter of rating, username, reviewId, review description, review date, reply and likes. Based on the given parameters, it will be made into column and exported into a csv format which can be seen from figure 5.

Fig. 5 CSV Data Format

Pre-Processing Data

Due to a lot of data had been crawl during the data scraping and need to be process. In addition, there are a lot of unstructured word like, short cut, emoticon, symbol, number and there are two language English and Indonesia in the data which required the method of pre-processing with the purpose to extract the necessary information from the review and change it into a standard word. In order to have an efficient workflow, author drop several irrelevant columns from original data table 4 to table 5.

Table 5. Original Data

	reviewtitle	review_description	rating	Thumbs_up	review_date	developer_response	developer_response_date	app Version	laguage_code	country_code
0	NaN	Good app	5	0	15/10/23 16:22	NaN	NaN	02.55.00	en	us
1	NaN	love it so far	4	0	15/10/23 12:38	NaN	NaN	02.55.00	en	us
2	NaN	super fast and free	5	0	15/10/23 10:18	NaN	NaN	02.54.00	en	us
3	NaN	okkkkkkkk	5	0	14/10/23 19:26	NaN	NaN	02.55.00	en	us
4	NaN	Nice n Love It	5	0	14/10/23 10:55	NaN	NaN	NaN	en	us
…	…	…	…	…	…	…	…	…	…	…
4895		Ok			02/07/21 16:19		02/07/21 16:41	2.02	en	us
4896		Mantaaap, gak sampai	5	2	02/07/21 15:00	Hai Sobat SeaBan	02/07/21 15:19	2.02	en	us

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13141
e-ISSN : 2541-2019
p-ISSN : 2541-044X

	5 menit, sudah punya rekening				k, terima kasih untuk reviewn ya …				
4897	Easy to use and simple	5	3	26/06/21 23:39	Hai Sobat SeaBank, terima kasih untuk reviewn ya …	27/06/27 12:44	2.01	en	us
4898	Nice app , mantab dan simpel	5	2	10/06/21 19:00	Hai Sobat SeaBank, terima kasih untuk reviewn ya …	11/06/21 12:05	NaN	en	us
4899	Very user friendly 👏	5	6	02/06/21 16:39	Hai Sobat SeaBank, terima kasih untuk reviewn ya …	07/06/21 19:17	1.5.0	en	us

Table 6. First data cleansing

	Review_description	rating
0	Good app	5
1	Love it so far	4
2	Super fast and free	5
3	okkkkkkkk	5
4	Nice n Love It	5
…	…	…
4895	Ok	5
4896	Mantaaap, gak sampai 5 menit, sudah punya rekening	5
4897	Easy to use and simple	5
4898	Nice app , mantab dan simpel	5
4899	Very user friendly 👏	5

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13141
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Case Folding and Labeling

Case folding is a step for converting review of capital or lower-case word or letter into a lower-case text. This process has been done to make sure that the word structure is in the same line. At the same time, author also put labeling per the given rating where rating 1,2,3 will be labelled as 0 (negative) and rating 4 & 5 is labelled as 1 (positive). The following table 6 is the result as for the rating graphic refer to figure 6.

Table 7. Case folding and Labeling.

	review_description	rating	lowcase	label
0	Good app	5	good app	1
1	love it so far	4	love it so far	1
2	super fast and free	5	super fast and free	1
3	okkkkkkkk	5	okkkkkkkk	1
4	Nice n Love It	5	nice n love it	1
…	…	…	…	…
4895	Ok	5	ok	1
4896	Mantaaap, gak sampai 5 menit, sudah punya rekening	5	mantaaap gak sampai menit sudah punya rekening	1
4897	Easy to use and simple	5	easy to use and simple	1
4898	Nice app , mantab dan simpel	5	nice app mantab dan simpel	1
4899	Very user friendly 👏	5	very user friendly	1

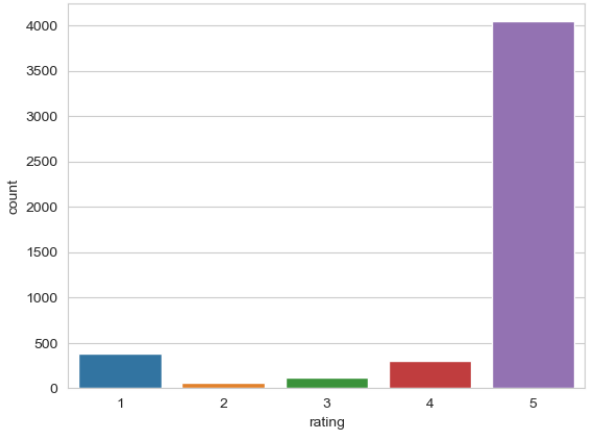

Fig. 6 Rating per labeling data

Tokenizing, Stop words & Lemmatization.

Tokenizing is the process of sentence become words, and after that we are going to the stage of stop words process of eliminating common words that don't have any purpose for example conjunction, adverbial word and etc., that don't have any impact toward the data analysis. For the stop words process, author is removing word based of stop word dictionary and put extension for Indonesian language common words as well. As for the lemmatizing process, it is changing the words of tokenize result into a base word. Following table 8 is the result of stop words and lemmatizing, refer to column lemmatized words.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13141
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 8. Tokenizing, Stop words and Lemmatization result

	review_ description	rating	low case	label	tokens	Lemmatized_ words
0	Good app	5	good app	1	[good, app]	good app
1	love it so far	4	love it so far	1	[love, it, so, far]	love far
2	super fast and free	5	super fast and free	1	[super, fast, and, free]	super fast free
3	okkkkkkkk	5	okkkkkkkk	1	[okkkkkkkk]	okkkkkkkk
4	Nice n Love It	5	nice n love it	1	[nice, n, love, it]	nice n love
…	…	…	…	…	…	…
4895	Ok	5	ok	1	[]	
4896	Mantaaap, gak sampai 5 menit, sudah punya rekening	5	mantaaap gak sampai menit sudah punya rekening	1	[mantaaap, menit, rekening]	mantaaap menit rekening
4897	Easy to use and simple	5	easy to use and simple	1	[easy, to, use, and, simple]	easy use simple
4898	Nice app , mantab dan simpel	5	nice app mantab dan simpel	1	[nice, app, mantab, simpel]	nice app mantab simpel
4899	Very user friendly 👆	5	very user friendly	1	[very, user, friendly]	user friendly

Data Training and Data Set

During the process of data training and data set, author is going to split the data into 5 cycles refer to table 9 for the results.

Table 9. Data Split of Training and Test

No.	Data Training		Data Testing	
	Percentage	Total Data	Percentage	Total Data
1	90%	4410	10%	490
2	70%	3430	30%	1470
3	50%	2450	50%	2450
4	30%	1470	70%	3430
5	10%	490	90%	4410

TF-IDF Process

After the splitting of the data, author is going to develop data feature for counting the value of every term with the method of TF-IDF (Term Frequency-Inverse Document Frequency) for producing value of the words that's previously extracted at lemmatized review of the train and test data. Text documents are converted using TF-IDF into a format that can be utilized for the NBC & SVM classification procedure (Lubis, Nasution, Sitompul, & Zamzami, 2021). For this process author is using the library sckit-learn with the module Tf-idf Vectorizer and Count Vectorizer.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13141
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 10. TF-IDF Data

No.	Data Training		Data Testing		TF-IDF Value
	Percentage	Total Data	Percentage	Total Data	
1	90%	4410	10%	490	1642
2	70%	3430	30%	1470	1367
3	50%	2450	50%	2450	1105
4	30%	1470	70%	3430	716
5	10%	490	90%	4410	291

Confusion Matrix

After splitting the data training and data testing, author will calculate the confusion matrix for evaluating the accuracy of classification model of NBC and SVM. Below table 11 and table 12 are the confusion matrix results for NBC and SVM based on 5 scenarios.

Table 11. Confusion Matrix NBC result

Confusion Matrix Naïve Bayes Classification – X_Test						
No.	Data Training	Data Testing	TP	FP	FN	TN
1	90%	10%	25	36	6	423
2	70%	30%	70	99	11	1290
3	50%	50%	105	181	18	2146
4	30%	70%	111	290	15	3014
5	10%	90%	25	483	2	3900

Table 12. Confusion Matrix SVM result

Confusion Matrix Support Vector Machine – X_Test						
No.	Data Training	Data Testing	TP	FP	FN	TN
1	90%	10%	31	30	13	416
2	70%	30%	94	75	23	1278
3	50%	50%	150	136	32	2132
4	30%	70%	160	241	30	2999
5	10%	90%	80	428	16	3886

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13141
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Classification Modeling by NBC and SVM

At this step, author is using the 5 scenarios of the data training and data testing for NBC and SVM algorithm by using the K-fold cross validation for evaluating the predictive models with the K = 10. Below is the evaluation result of each scenario with parameters of Accuracy, Precision, recall and F1-Score.

Table 13. NBC K=10

Naïve Bayes Classification by K=10						
No.	**Data Training**	**Data Testing**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
1	90%	10%	93.60%	93.96%	99.18%	96.50%
2	70%	30%	93.11%	93.44%	99.21%	96.24%
3	50%	50%	92.89%	93.20%	99.26%	96.14%
4	30%	70%	92.44%	92.84%	99.24%	95.93%
5	10%	90%	91.63%	91.59%	100%	95.60%

Table 14. SVM K=10

(SVM) Support Vector Machine by K=10						
No.	**Data Training**	**Data Testing**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
1	90%	10%	93.99%	94.60%	98.87%	96.69%
2	70%	30%	93.46%	94.04%	98.91%	96.41%
3	50%	50%	93.67%	93.98%	99.26%	96.55%
4	30%	70%	92.85%	93.36%	99.08%	96.13%
5	10%	90%	91.42%	91.56%	99.77%	95.49%

Sentiment Analysis Review

Based on above model classification, the best performance is on scenario one with the data training of 90% and data test of 10%, with the algorithm of SVM. Below is the word cloud visualization of the top used words in positive and negative review.

Fig. 7 Word Cloud Positive and Negative result

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13141
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Website Implementation

Based on the best sentiment analysis algorithm, author is developing a sentiment website by using flask framework and python programming code. And the design is created using HTML and CSS, with 2 pages which are home and result. Hereafters are the final look of the website. Author had also deploy the website to Heroku App can be checked on this link https://sentiment-digital-bank-0e85e52c0405.herokuapp.com/

a. Home Page
In this home page user can input their desired comments related to digital bank for checking whether this review fall into negative or positive section.

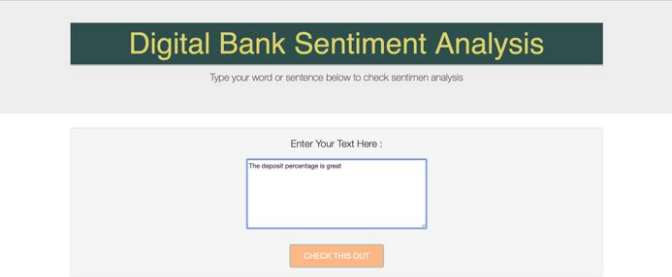

Fig. 8 Home Page

b. Result page
Once an input had been done in Homepage, there will be 2 types of result which are positive review and negative review, therefore, user can have more understanding if it's a good or bad result.

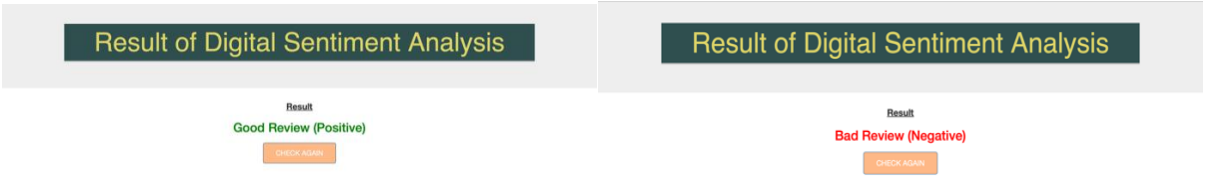

Fig. 9 Positive and Negative result

DISCUSSIONS

The author used the Naïve Bayes and Support Vector Machine algorithms to assess the Google Play Store review of SEA Bank, a digital bank. Understanding the purpose of the study, its methodology, and the desired outcomes is the first step in the process. in order to fulfill the necessary requirements. Prior to moving on to the next phase, there will be a data cleansing (pre-processing) phase. The author in this instance is determining the accuracy, precision, recall, and F1-Score using K-fold cross validation (K=10). additionally depending on the parameter comparison between SVM and NBC. Because the data indicates that SVM is a superior algorithm than others, the author decided to move forward with the development of the app.

If we compare the present research with other research that is referenced in the literature review, for instance, a research publication titled "analysis sentiment toward WhatsApp new privacy and policy for Naïve Bayes Classification and Support Vector Machine". The studies concentrate on the analysis approach and solely use the confusion matrix and accuracy to establish the outcome, which shows 338 favorable and 422 unfavorable evaluations regarding the new WhatsApp policy.

This research study's weakness is the small amount of Google Play review data that was used to precisely identify positive and negative reviews when they were provided for a web application.

*name of corresponding author

CONCLUSION

In this study, author is gathering data with the Google Play Scraper Python Module to obtain the study case sea bank data from January 2021 to October 2023, totaling 4899 data points. This study employs a research and development methodology. Specifically, the author uses both the NBC and SVM algorithms to analyze data and determine which classification model performs best. The results show that splitting the data into 90% for training and 10% for testing produced the best results in the experiments. An F1 score of 96.69% and accuracy, precision, and recall rates of 93.99%, 94.60%, and 98.87%, respectively, were obtained in this scenario with a k-value of 10.

These results were obtained through the utilization of the Support Vector Machine (SVM) classification algorithm.

After the author has determined which classification model works best, the next stage is development using waterfall methodology with flask framework. By developing a sentiment website that user can input their feedback and determine whether its favorable or unfavorable.

REFERENCES

Aditra Pradnyana, G., Gede, I., & Darmawiguna, M. (2021). *Web-Based System for Bali Tourism Sentiment Analysis during The Covid-19 Pandemic using Django Web Framework and Naive Bayes Method*.

Alawi, S. J. S. Al, Jamil Jastini Mohd, & Shaharanee, I. N. M. (2022). Predicting Student Performance Using Data Mining Approach: A Case Study in Oman. *Publication Issue*, *71*(4), 1389–1398. Retrieved from http://philstat.org.ph

Cahyaningrum, A. D., & Atahau, A. D. R. (2021). Intellectual Capital And Financial Performance: Banks' Risk As The Mediating Variable. *Jurnal Manajemen Dan Kewirausahaan*, *22*(1), 21–32. https://doi.org/10.9744/jmk.22.1.21-32

Cupian, C., & Akbar, F. F. (2020). Analisis Perbedaan Tingkat Profitabilitas Perbankan Syariah Sebelum Dan Setelah Bekerja Sama Dengan Perusahaan Financial Technology (Fintech) (Studi Kasus Bank Bni Syariah, Bank Syariah Mandiri, Dan Bank Mega Syariah). *Jurnal Ekonomi Syariah Teori Dan Terapan*, *7*(11), 2149. https://doi.org/10.20473/vol7iss202011pp2149-2169

Fransiska, S., & Irham Gufroni, A. (2020). Sentiment Analysis Provider by.U on Google Play Store Reviews with TF-IDF and Support Vector Machine (SVM) Method. *Scientific Journal of Informatics*, *7*(2), 2407–7658. Retrieved from http://journal.unnes.ac.id/nju/index.php/sji

Ilmawan, L. B., & Mude, M. A. (2020). Perbandingan Metode Klasifikasi Support Vector Machine dan Naïve Bayes untuk Analisis Sentimen pada Ulasan Tekstual di Google Play Store. *ILKOM Jurnal Ilmiah*, *12*(2), 154–161. https://doi.org/10.33096/ilkom.v12i2.597.154-161

Lubis, A. R., Nasution, M. K. M., Sitompul, O. S., & Zamzami, E. M. (2021). The effect of the TF-IDF algorithm in times series in forecasting word on social media. *Indonesian Journal of Electrical Engineering and Computer Science*, *22*(2), 976–984. https://doi.org/10.11591/ijeecs.v22.i2.pp976-984

Maulana, A. A., Susanto, A., & Kusumaningrum, D. P. (2019). Rancang Bangun Web Scraping Pada Marketplace di Indonesia. *JOINS (Journal of Information System)*, *4*(1), 41–53. https://doi.org/10.33633/joins.v4i1.2544

Nur'aini, A., & Alfirman. (2021). *Analisa Sentimen Pengguna terhadap kebijakan baru whatsapp menggunakan Naive Bayes Classifier dan Support Vector Machine*. 1–14. Retrieved from https://www.ptonline.com/articles/how-to-get-better-mfi-results

Rudra Kumar, M., Pathak, R., & Gunjan, V. K. (2022). Diagnosis and Medicine Prediction for COVID-19 Using Machine Learning Approach. *Lecture Notes in Electrical Engineering*, *834*, 123–133. Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-981-16-8484-5_10

Salehudin Basryah, E., Erfina, A., & Warman, C. (2021). *Analisis Sentimen Aplikasi Dompet Digital Di Era 4.0 Pada Masa Pendemi Covid-19 Di Play Store Menggunakan Algoritma Naive Bayes Classifier*. 189–196.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i1.13141 p-ISSN : 2541-044X

Shamantha Rai B, & Shetty Sweekriti M. (2019). *Sentiment Analysis Using Machine Learning Classifiers: Evaluation of Performance*.

Wahyu Handani, S., Intan Surya Saputra, D., Hasirun, Mega Arino, R., & Fiza Asyrofi Ramadhan, G. (2019). Sentiment analysis for go-jek on google play store. *Journal of Physics: Conference Series*, *1196*(1). Institute of Physics Publishing. https://doi.org/10.1088/1742-6596/1196/1/012032

Ridhoi, M. A. (2021). Selamat Datang Era Bank Digital di Indonesia, Prospek & Tantangannya. https://katadata.co.id/muhammadridhoi/analisisdata/5fe2d448aca0a/selamat-datang-era-bank-digital-di indonesia-prospek-tantangannya

Ramli, R. R. (2021). Bank Digital Terus Tumbuh, Ekonomi Digital Indonesia Diproyeksi Jadi Terbesar Se-Asia Tenggara pada 2025. https://money.kompas.com/read/2021/07/01/175703526/bank-digital-terus-tumbuh-ekonomi-digital-indonesia-diproyeksi-jadi-terbesar?page=all

*name of corresponding author