

Comparison of Support Vector Machine, Random Forest and XGBoost for Sentiment Analysis on Indodax

Moch. Alfarros Difa Naufalino^{1)*}, Rianto²⁾, Muhammad Al-Husaini³⁾

^{1,2,3)}Faculty of Engineering and Informatics Engineering, Siliwangi University, Tasikmalaya, Indonesia

¹⁾207006015@student.unsil.ac.id, ²⁾rianto@unsil.ac.id, ³⁾alhusaini@unsil.ac.id

ABSTRACT

The rapid growth of digital assets like Bitcoin and cryptocurrencies has increased the need for secure trading platforms such as Indodax. With the growing number of users, reviews on platforms like Google Play Store provide valuable insights into user experience and satisfaction. This research applies Machine Learning methods to classify user review sentiments by comparing three main algorithms Support Vector Machine (SVM), Random Forest, and Extreme Gradient Boosting (XGBoost). One of the main challenge in sentiment analysis is the presence of irrelevant or redundant features, which can reduce model accuracy and increase computational costs. The Feature Selection Chi-Square technique is used to filter the most influential features, enhancing model efficiency without losing critical information. Experimental results show that SVM delivers the best performance compared to Random Forest and XGBoost. Before applying Chi-Square, SVM achieved 91% accuracy, which increased to 94% after applying the feature selection technique. The number of features used was reduced from 52,312 to 2,000 without significant information loss. This combination of SVM and Feature Selection Chi-Square proves to be an efficient and accurate solution for analyzing user sentiment on crypto trading platforms like Indodax. This method is expected to improve the responsiveness of trading applications to user needs and serve as a foundation for further research in Machine Learning-based sentiment analysis.

Keywords: Feature Selection Chi-Square, Machine Learning, Sentiment Analysis, Support Vector Machine, Random Forest, XGBoost

INTRODUCTION

Information technology and the internet have provided many benefits in various aspects of life, including economics, trade and investment. Bitcoin and cryptocurrencies are forms of digital investment that are increasingly in demand despite not being recognized as legal tender. Along with the increasing public interest in digital assets, trading platforms play an important role in providing safe and easy access for users. The popularity of cryptocurrencies in Indonesia continues to increase, as shown by the growing number of users and transactions. This emphasizes the importance of user experience in using crypto trading applications. One way to understand user experience and satisfaction is through reviews provided on platforms such as the Google Play Store (Sunarko et al., 2023).

Indodax is one of the largest crypto asset trading platforms in Indonesia with 7.4 million registered users. User reviews given on the Google Play Store reflect their experiences, whether in the form of suggestions, criticisms, or complaints, which can be a reference for potential users (Permana et al., 2023). Sentiment analysis of these reviews can provide insights into the extent to which the app meets user expectations, service reliability, and technical issues that may occur. Machine learning technology can be used to automatically and effectively analyze reviews to get an overview of user satisfaction (Aida Sapitri et al., 2023).

In the sentiment analysis of Indodax app, various machine learning methods have been used, such as Support Vector Machine (SVM), Random Forest, and Extreme Gradient Boosting (XGBoost). SVM is used to find the best separating boundary among different sentiment classes (Oktavia et al., 2023). Previous research shows that the use of SVM in sentiment analysis of Grab app reviews achieved 85.54% accuracy (Wahyudi et al., 2021). Random Forest was used to reduce overfitting by combining multiple decision trees (Repina Kesuma et al., 2020), while another study recorded 67% accuracy on sentiment analysis related to the MUI Fatwa controversy (Yasir et al., n.d.). Meanwhile, XGBoost utilizes boosting techniques to improve model accuracy in sentiment classification, with previous research showing 71% accuracy on online shopping app review analysis (Sartina Jayanti et al., n.d.).

However, one of the challenges in sentiment analysis is the presence of redundant features, which can lead to a decrease in model efficiency (Erkamim et al., 2023). To overcome this problem, Feature Selection Chi-Square can be applied as a feature selection technique that is able to eliminate less informative features without significantly

* Corresponding author



[Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.](https://creativecommons.org/licenses/by-nc-sa/4.0/)

degrading model accuracy (Ernayanti et al., 2023). This technique aims to improve the efficiency and accuracy of the model in sentiment classification.

Based on previous research, there is an opportunity to develop a more optimal method by comparing SVM, Random Forest, and XGBoost after applying Feature Selection Chi-Square. With this approach, this research aims to identify the best method in the sentiment classification of Indodax app reviews and overcome the problem of redundant data (Ernayanti et al., 2023). Hopefully this research can contribute to the selection of the best algorithm by considering the number of features that contribute most to accuracy, so that the classification process can be carried out more efficiently and effectively.

LITERATURE REVIEW

Several studies have shown that machine learning algorithms such as Support Vector Machine, Random Forest, and XGBoost often face challenges when processing high-dimensional textual data, which can lead to irrelevant or redundant features (Erkamim et al., 2023) highlighted the limitations of using Support Vector Machine, Random Forest, and XGBoost algorithms in text data processing, especially regarding irrelevant or redundant results. These problems often occur due to too large feature dimensions without adequate feature selection techniques. This research is the basis that in text data processing such as sentiment analysis, additional approaches are needed to increase the relevance of features to make classification results more accurate and efficient.

(Tasnim & Habiba, 2021) showed that applying the Chi-Square feature selection method can significantly improve the performance of classification algorithms. In their study, applying Chi-Square to the Random Forest algorithm resulted in an accuracy of 92.85%, which is higher than using the algorithm without feature selection. This result strengthens the argument that Chi-Square is able to eliminate unimportant features, thus improving the quality of predictions in text analysis.

Research by (Mat et al., 2022) developed an Android malware detection system using permission features and Chi-Square and Information Gain feature selection methods. With a dataset of 10,000 samples from AndroZoo and Drebin, experiments showed that the Chi-Square-based classification model was able to achieve an accuracy rate of 91.1%. This study confirms that the Chi-Square method is effective not only in general text classification, but also in the domain of digital security and malware detection.

(Yudha et al., 2022) applied Chi-Square feature selection to the Convolutional Neural Network (CNN) model for the classification of Indonesian news articles into six categories. The results showed that the combination of CNN and Chi-Square produced an accuracy of up to 96.07%, with almost equal precision, recall, and F1-score values. The application of Chi-Square in the context of news classification demonstrates the method's flexibility in handling large amounts of complex text.

In a study conducted by (Chen Sami et al., 2024), sentiment analysis of data privacy issues on YouTube social media was conducted with the support of Chi-Square feature selection and C5.0 classification method. This research shows that the accuracy of the model increased from 80.14% to 87.34% after applying Chi-Square. This proves that feature selection not only improves accuracy, but also strengthens the relevance of the features to the sentiment topics studied, so that the analysis results become sharper.

Research by (Bhoria et al., 2023) proposed the Chi-Square-Based Feature Selection (CBFS) method in predicting student academic performance using data from UCI with 32 attributes. Models such as Decision Tree, Random Forest, SVM, KNN, and XGBoost were compared both with and without feature selection. As a result, SVM recorded the highest F1 score of 95.19%, and all models showed improved performance when using Chi-Square. This strengthens the understanding that statistical-based feature selection can improve model efficiency in various data contexts, including education.

(Afriyani et al., 2024) examined the effectiveness of Chi-Square feature selection in improving classification accuracy using SVM, KNN, and Random Forest on 2024 General Election public comment data. With the addition of Pseudo-Labeling technique for semi-supervised learning, the accuracy reached 92%, showing that the combination of Chi-Square and Pseudo-Labeling can improve classification performance in NLP, especially when labelled data is limited.

METHOD

The stages of the research flow include dataset collection, preprocessing, labeling using lexicon, word cloud creation, modeling with SVM, Random Forest, and XGBoost, and presentation of modeling results using confusion matrix. The research flow can be seen in Figure 1.

* Corresponding author



[Creative Commons Attribution-NonCommercial-ShareAlike 4.0
International License.](https://creativecommons.org/licenses/by-nc-sa/4.0/)

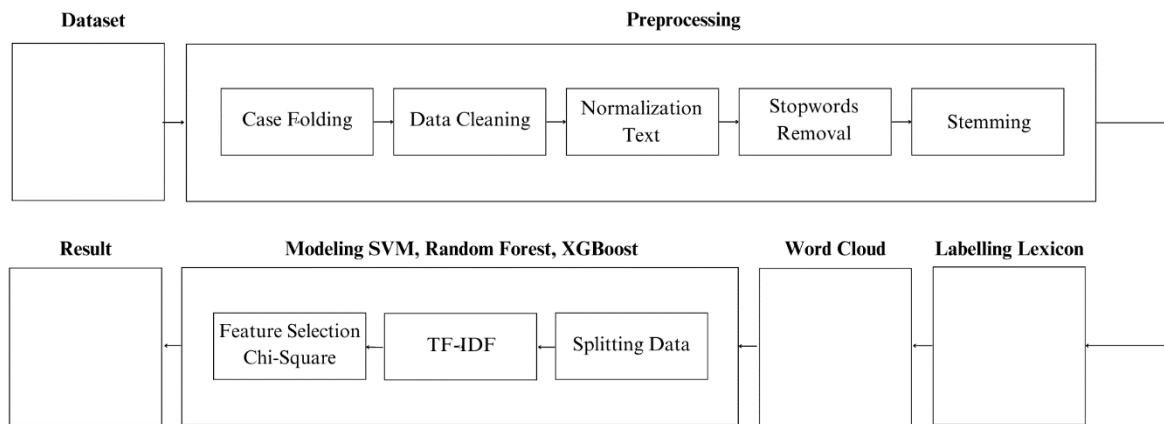


Fig. 1 Research Stages

Scraping Dataset

The first stage in this research is to collect the comment dataset in .csv format. The dataset was obtained using the scraping technique on the play store application for the Indodax application. The scraping process can be seen in Figure 2.

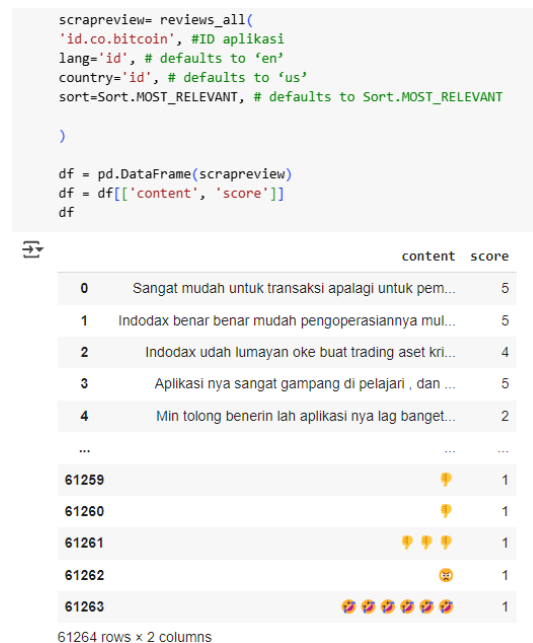


Fig 2. Dataset Scraping Process

Figure 2 is dataset scraping process using the 'google_play_scraper' library to retrieve app review data from the Google Play Store. The 'reviews_all()' function is used to retrieve all reviews of an app with a specific ID ('id.co.bitcoin'). The reviews are filtered by language ('id' for Bahasa Indonesia) and country ('id' for Indonesia). The data is then stored in a Pandas DataFrame, retrieving only the 'content' and 'score' columns, which can later be used for the modeling process.

Preprocessing

After the dataset has been successfully obtained through the scraping process, the next step is the preprocessing which aims to clean and prepare data to make it more qualified before being used in the model training process

* Corresponding author



[Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.](https://creativecommons.org/licenses/by-nc-sa/4.0/)

(Rivanie et al., 2021). The first process is case folding, can be seen in Table 1.

Table 1. Case Folding

content	casefolding
Indodax udah lumayan oke buat trading aset kripto, terutama untuk pemula. UI-nya gampang dipahami, banyak pilihan koin, sama proses deposit/withdrawal yang lumayan cepat. Suka juga karena ada fitur edukasi buat yang baru belajar. Tapi kekurangannya, kadang sistemnya suka delay pas market lagi rame, dan fee tradingnya lumayan kerasa kalau sering transaksi. Semoga ke depannya bisa lebih stabil dan ada promo fee yang lebih kompetitif. Username: Sulton_Hasanudin #AsetMasaDepan	indodax udah lumayan oke buat trading aset kripto, terutama untuk pemula. ui-nya gampang dipahami, banyak pilihan koin, sama proses deposit/withdrawal yang lumayan cepat. suka juga karena ada fitur edukasi buat yang baru belajar. tapi kekurangannya, kadang sistemnya suka delay pas market lagi rame, dan fee tradingnya lumayan kerasa kalau sering transaksi. semoga ke depannya bisa lebih stabil dan ada promo fee yang lebih kompetitif. username: sulton_hasanudin #asetmasadepan
Aplikasi nya sangat gampang di pelajari , dan tampilan nya yg minimalis tetapi detail. Membuat enak di lihat. Semoga aplikasi indodax ini selalu sukses dan pemeliharaan nya selalu di perhatikan agar orang orang yg baru bergabung di kripto mudah memahami	aplikasi nya sangat gampang di pelajari , dan tampilan nya yg minimalis tetapi detail. Membuat enak di lihat. Semoga aplikasi indodax ini selalu sukses dan pemeliharaan nya selalu di perhatikan agar orang orang yg baru bergabung di kripto mudah memahami
Min tolong benerin lah aplikasi nya lag banget, server nya lemot, suka keluar sendiri, buka chart sering error. Ga cuman saya yg ngalamin di chat room indodax banyak yg merasakan hal yg sama juga. Saya sudah update aplikasi nya tetep aja udah uninstall pasang ulang juga tetep lag	min tolong benerin lah aplikasi nya lag banget, server nya lemot, suka keluar sendiri, buka chart sering error. Ga cuman saya yg ngalamin di chat room indodax banyak yg merasakan hal yg sama juga. Saya sudah update aplikasi nya tetep aja udah uninstall pasang ulang juga tetep lag

Table 1 is the case folding process to standardize the writing format, which initially varies between uppercase and lowercase letters to all lowercase letters. Case folding is used to get more efficient features with the same format. The next stage is data cleaning, can be seen in Table 2.

Table 2. Data Cleaning

casefolding	cleaning
indodax udah lumayan oke buat trading aset kripto, terutama untuk pemula. ui-nya gampang dipahami, banyak pilihan koin, sama proses deposit/withdrawal yang lumayan cepat. suka juga karena ada fitur edukasi buat yang baru belajar. tapi kekurangannya, kadang sistemnya suka delay pas market lagi rame, dan fee tradingnya lumayan kerasa kalau sering transaksi. semoga ke depannya bisa lebih stabil dan ada promo fee yang lebih kompetitif. username: sulton hasanudin #asetmasadepan	indodax udah lumayan oke buat trading aset kripto terutama untuk pemula ui nya gampang dipahami banyak pilihan koin sama proses deposit withdrawal yang lumayan cepat suka juga karena ada fitur edukasi buat yang baru belajar tapi kekurangannya kadang sistemnya suka delay pas market lagi rame dan fee tradingnya lumayan kerasa kalau sering transaksi semoga ke depannya bisa lebih stabil dan ada promo fee yang lebih kompetitif username sulton hasanudin asetmasadepan
aplikasi nya sangat gampang di pelajari , dan tampilan nya yg minimalis tetapi detail. Membuat enak di lihat. Semoga aplikasi indodax ini selalu sukses dan pemeliharaan nya selalu di perhatikan agar orang orang yg baru bergabung di kripto mudah memahami	aplikasi nya sangat gampang di pelajari dan tampilan nya yg minimalis tetapi detail membuat enak di lihat semoga aplikasi indodax ini selalu sukses dan pemeliharaan nya selalu di perhatikan agar orang orang yg baru bergabung di kripto mudah memahami
min tolong benerin lah aplikasi nya lag banget, server nya lemot, suka keluar sendiri, buka	min tolong benerin lah aplikasi nya lag banget server nya lemot suka keluar sendiri buka chart

* Corresponding author



[Creative Commons Attribution-NonCommercial-ShareAlike 4.0
International License.](https://creativecommons.org/licenses/by-nc-sa/4.0/)

chart sering error. Ga cuman saya yg ngalamin di chat room indodax banyak yg merasakan hal yg sama juga. Saya sudah update aplikasi nya tetep aja udah uninstall pasang ulang juga tetep lag	sering error ga cuman saya yg ngalamin di chat room indodax banyak yg merasakan hal yg sama juga saya sudah update aplikasi nya tetep aja udah uninstall pasang ulang juga tetep lag
--	--

Table 2. is the data cleaning process of cleaning review data that has symbols or characters other than alphabets (Nurtikasari et al., 2022). The next stage is normalization text, can be seen in Table 3.

Table 3. Normalization Text

cleaning	textnormalize
indodax udah lumayan oke buat trading aset kripto terutama untuk pemula ui nya gampang dipahami banyak pilihan koin sama proses deposit withdrawal yang lumayan cepat suka juga karena ada fitur edukasi buat yang baru belajar tapi kekurangannya kadang sistemnya suka delay pas market lagi rame dan fee tradingnya lumayan kerasa kalau sering transaksi semoga ke depannya bisa lebih stabil dan ada promo fee yang lebih kompetitif username sulton hasanudin asetmasadepan	indodax sudah lumayan oke buat trading aset kripto terutama untuk pemula ui nya gampang dipahami banyak pilihan koin sama proses deposit withdrawal yang lumayan cepat suka juga karena ada fitur edukasi buat yang baru belajar tapi kekurangannya kadang sistemnya suka delay pas market lagi ramai dan biaya tradingnya lumayan terasa kalau sering transaksi semoga ke depannya bisa lebih stabil dan ada promo biaya yang lebih kompetitif username sulton hasanudin asetmasadepan
aplikasi nya sangat gampang di pelajari dan tampilan nya yg minimalis tetapi detail membuat enak di lihat semoga aplikasi indodax ini selalu sukses dan pemeliharaan nya selalu di perhatikan agar orang orang yg baru bergabung di kripto mudah memahami	aplikasi nya sangat gampang di pelajari dan tampilan nya yang minimalis tetapi detail membuat enak di lihat semoga aplikasi indodax ini selalu sukses dan pemeliharaan nya selalu di perhatikan agar orang orang yang baru bergabung di kripto mudah memahami
min tolong benerin lah aplikasi nya lag banget server nya lemot suka keluar sendiri buka chart sering error ga cuman saya yg ngalamin di chat room indodax banyak yg merasakan hal yg sama juga saya sudah update aplikasi nya tetep aja udah uninstall pasang ulang juga tetep lag	min tolong benerin lah aplikasi nya lag banget server nya lemot suka keluar sendiri buka chart sering error tidak Cuma saya yang ngalamin di chat room indodax banyak yang merasakan hal yang sama juga saya sudah update aplikasi nya tetap saja sudah uninstall pasang ulang juga tetap lag

Table 3 is the process of normalization text is informal or non-standard words into a more standard form. The next stage is stopwords removal, can be seen in Table 4.

Table 4. Stopwords Removal

textnormalize	stopwordremoval
indodax sudah lumayan oke buat trading aset kripto terutama untuk pemula ui nya gampang dipahami banyak pilihan koin sama proses deposit withdrawal yang lumayan cepat suka juga karena ada fitur edukasi buat yang baru belajar tapi kekurangannya kadang sistemnya suka delay pas market lagi ramai dan biaya tradingnya lumayan terasa kalau sering transaksi semoga ke depannya bisa lebih stabil dan ada promo biaya yang lebih kompetitif username sulton hasanudin asetmasadepan	indodax lumayan oke trading aset kripto pemula ui nya gampang dipahami pilihan koin proses deposit withdrawal lumayan cepat suka fitur edukasi belajar kekurangannya kadang sistemnya suka delay pas market ramai biaya tradingnya lumayan transaksi semoga depannya stabil promo biaya kompetitif username sulton hasanudin asetmasadepan

aplikasi nya sangat gampang di pelajari dan tampilan nya yang minimalis tetapi detail membuat enak di lihat semoga aplikasi indodax ini selalu sukses dan pemeliharaan nya selalu di perhatikan agar orang orang yang baru bergabung di krypto mudah memahami	aplikasi nya gampang pelajari tampilan nya minimalis detail enak lihat semoga aplikasi indodax sukses pemeliharaan nya perhatikan orang orang bergabung krypto mudah memahami
min tolong benerin lah aplikasi nya lag banget server nya lemot suka keluar sendiri buka chart sering error tidak cuma saya yang ngalamin di chat room indodax banyak yang merasakan hal yang sama juga saya sudah update aplikasi nya tetap saja sudah uninstall pasang ulang juga tetap lag	min tolong benerin aplikasi nya lag banget server nya lemot suka buka chart error ngalamin chat room indodax merasakan update aplikasi nya uninstall pasang ulang lag

Table 4 is the process of stopwords removal, which are words/features that have no meaning/relationship (irrelevant) to the main subject (Pasek et al., 2022). Stopwords removal is a step to remove features that have no meaning and no impact on the text analysis process. The next stage is stemming, can be seen in Table 5.

Table 5. Stemming

stopwordremoval	stemming
indodax lumayan oke trading aset krypto pemula ui nya gampang dipahami pilihan koin proses deposit withdrawal lumayan cepat suka fitur edukasi belajar kekurangannya kadang sistemnya suka delay pas market ramai biaya tradingnya lumayan transaksi semoga depannya stabil promo biaya kompetitif username sulton hasanudin asetmasadepan	indodax lumayan oke trading aset krypto mula ui nya gampang paham pilih koin proses deposit withdrawal lumayan cepat suka fitur edukasi ajar kurang kadang sistem suka delay pas market ramai biaya tradingnya lumayan transaksi moga depan stabil promo biaya kompetitif username sulton hasanudin asetmasadepan
aplikasi nya gampang pelajari tampilan nya minimalis detail enak lihat semoga aplikasi indodax sukses pemeliharaan nya perhatikan orang orang bergabung krypto mudah memahami	aplikasi nya gampang ajar tampil nya minimal detail enak lihat moga aplikasi indodax sukses pemeliharaan nya perhati orang orang gabung krypto mudah paham
min tolong benerin aplikasi nya lag banget server nya lemot suka buka chart error ngalamin chat room indodax merasakan update aplikasi nya uninstall pasang ulang lag	min tolong benerin aplikasi nya lag banget server nya lemot suka buka chart error ngalamin chat room indodax rasa update aplikasi nya uninstall pasang ulang lag

Table 5 is the stemming process, which returns the words obtained from the filtering results to their basic form (Petronella Purba & Transver Wijaya, n.d.).

Labelling Lexicon

The next stage is labelling lexicon, which is a method of labeling sentiment on text based on a word dictionary (lexicon) containing a list of positive and negative words along with their sentiment weights.

Table 6. Labelling Lexicon

content	clean text	sentiment	weight
Indodax udah lumayan oke buat trading aset krypto, terutama untuk pemula. UI-nya gampang dipahami, banyak pilihan koin, sama proses deposit/withdrawal yang lumayan cepat. Suka juga karena ada fitur edukasi buat yang baru belajar. Tapi kekurangannya, kadang sistemnya suka delay pas market lagi rame, dan fee	indodax lumayan oke trading aset krypto mula ui nya gampang paham pilih koin proses deposit withdrawal lumayan cepat suka fitur edukasi ajar kurang kadang sistem suka delay pas market ramai biaya tradingnya lumayan transaksi moga depan stabil promo biaya kompetitif username sulton hasanudin asetmasadepan	positive	1

* Corresponding author



[Creative Commons Attribution-NonCommercial-ShareAlike 4.0
International License.](https://creativecommons.org/licenses/by-nc-sa/4.0/)

content	clean_text	sentiment	weight
tradingnya lumayan kerasa kalau sering transaksi. Semoga ke depannya bisa lebih stabil dan ada promo fee yang lebih kompetitif. Username: Sulton_Hasanudin #AsetMasaDepan			
Aplikasi nya sangat gampang di pelajari , dan tampilan nya yg minimalis tetapi detail. Membuat enak di lihat. Semoga aplikasi indodax ini selalu sukses dan pemeliharaan nya selalu di perhatikan agar orang orang yg baru bergabung di kripto mudah memahami	aplikasi nya gampang ajar tampil nya minimal detail enak lihat moga aplikasi indodax sukses pemeliharaan nya perhati orang orang gabung kripto mudah paham	positive	1
Min tolong benerin lah aplikasi nya lag banget, server nya lemot, suka keluar sendiri, buka chart sering error. Ga cuman saya yg ngalamin di chat room indodax banyak yg merasakan hal yg sama juga. saya sudah update aplikasi nya tetep aja udah uninstall pasang ulang juga tetep lag	min tolong benerin aplikasi nya lag banget server nya lemot suka buka chart error ngalamin chat room indodax rasa update aplikasi nya uninstall pasang ulang lag	negative	-1

Word Cloud

The next stage is Word Cloud which functions to display the words that appear most often in the text based on sentiment (positive and negative) in visual form. The result of labelling lexicon can be seen in Figure 3.

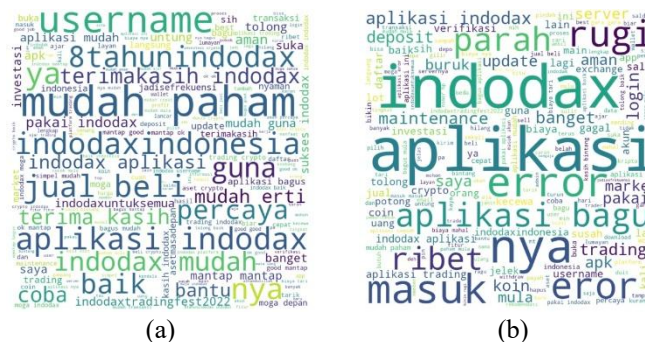


Fig 3. (a) Word Cloud Positive Sentiment (b) World Cloud Negative Sentiment

Modeling

The next stage is modeling using SVM, Random Forest and XGBoost. Modeling is done by splitting the data first using a ratio of 80:20 for test and training (N. B. Putri & Wijayanto, 2022), and then TF-IDF performs the process of converting text into numerical representation so that Feature Selection Chi-Square can be used. Feature Selection Chi-Square serves to eliminate some irrelevant features. The features taken are 2,000 features from 52.312 based on the highest Chi-Square value which indicates that the features have a significant correlation with the sentiment label.

Evaluation

The next step is to review the results of the modeling using confusion matrix and perform a comparison of each algorithm used to determine the one that produces the most optimal accuracy.

a) Accuracy

The proportion of correct predictions from all predictions made. The accuracy calculation equation can be seen in the equation (1)

* Corresponding author



[Creative Commons Attribution-NonCommercial-ShareAlike 4.0
International License.](https://creativecommons.org/licenses/by-nc-sa/4.0/)

$$\frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (1)$$

b) Precision

The proportion of correct positive predictions from all positive predictions. The precision calculation equation can be seen in the equation (2)

$$\frac{TP}{TP + FP} \times 100\% \quad (2)$$

c) Recall

The proportion of all positive cases detected correctly by the model. The recall calculation equation can be seen in equation (3)

$$\frac{TP}{TP + FN} \times 100 \quad (3)$$

d) F1-Score

The harmonic mean of precision and recall, gives a balanced picture between the two. The F1-Score calculation equation can be seen in the equation (4)

$$2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (4)$$

where :

True Positive (TP) is the number of correct predictions for the positive class, where the model correctly identifies the positive sentiment.

True Negative (TN) is the number of correct predictions for the negative class, where the model correctly identifies the negative sentiment.

False Positive (FP) is the number of false predictions for the positive class, where the model classifies the negative sentiment as positive.

False Negative (FN) is the number of false predictions for the negative class, where the model classifies positive sentiment as negative.

RESULT

The results of model training using SVM, Random Forest, and XGBoost are obtained through the calculation of the confusion matrix for each model, which is then used to compare model performance. Modeling is divided into two parts, namely without Feature Selection Chi-Square and with Feature Selection Chi-Square. Comparison of model results is done to evaluate performance before and after the application of Feature Selection Chi-Square. The training results of SVM, Random Forest, and XGBoost models without Feature Selection Chi-Square can be seen in Figure 4.

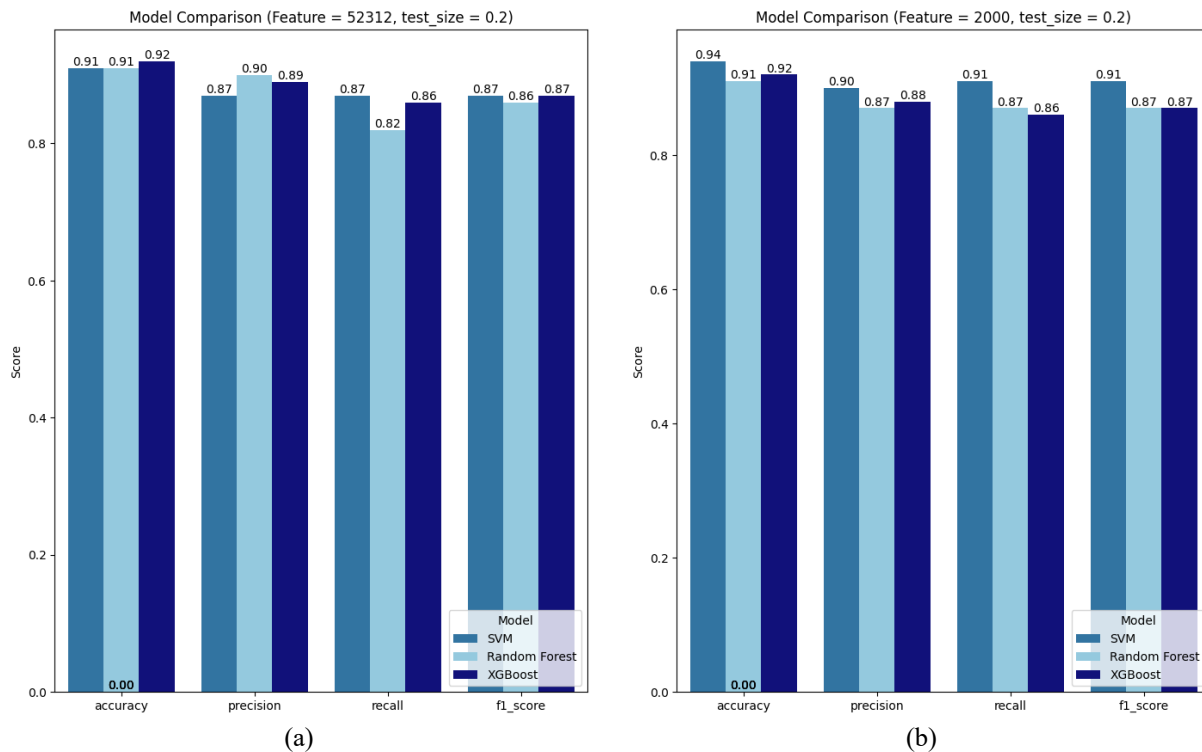


Fig 4. Model Comparison (a) without Feature Selection Chi-Square (b) with Feature Selection Chi-Square

Figure 4 shows a bar graph comparing the performance of the three models, namely SVM, Random Forest, and XGBoost models in graph (a) without Feature Selection Chi-Square and graph (b) after Feature Selection Chi-Square. In graph (a), the model works with 52.312 features and produces highest accuracy 0.92 in XGBoost. After feature selection in graph (b) by reducing the number of features to 2.000, the performance of the model improves. SVM showed an increase in accuracy from 0.91 to 0.94, while Random Forest and XGBoost also saw improvements in precision, recall and F1-score. This proves that Feature Selection Chi-Square can improve model efficiency without sacrificing accuracy. The improvement in the value of each model can be seen in table 7.

Table 7. Model Performance Before and After Feature Selection Chi-Square

Model	Metric	Without Chi-Square	With Chi-Square	Improvement
SVM	Accuracy	0,91	0,94	+0,03
	Precision	0,87	0,90	+0,03
	Recall	0,87	0,91	+0,04
	F1-Score	0,87	0,91	+0,04
Random Forest	Accuracy	0,91	0,91	-
	Precision	0,90	0,87	-
	Recall	0,82	0,87	+0,05
	F1-Score	0,86	0,87	+0,01
XGBoost	Accuracy	0,92	0,92	-
	Precision	0,89	0,88	-
	Recall	0,86	0,86	-
	F1-Score	0,87	0,87	-

Based on the performance comparison table of Machine Learning models (SVM, Random Forest, and XGBoost) before and after using Feature Selection Chi-Square, it can be seen that the Chi-Square method has the most significant impact on the SVM model with improvements in all performance metrics, where accuracy increases from 0.91 to 0.94, precision from 0.87 to 0.90, recall from 0.87 to 0.91, and F1-score from 0.87 to 0.91. In the Random Forest model,

Chi-Square only increased recall from 0.82 to 0.87 and F1-score from 0.86 to 0.87, while accuracy and precision remained unchanged. In contrast, in the XGBoost model Chi-Square had no significant effect because all performance metrics (accuracy, precision, recall, and F1-score) remained the same before and after the application of Chi-Square. This shows that Chi-Square is effective in improving the performance of SVM models, helpful in Random Forest models, but has no significant impact on XGBoost models.

DISCUSSIONS

This research shows good results from using Feature Selection Chi-Square in sentiment analysis models. Support Vector Machine, Random Forest, and XGBoost models, experienced an increase in accuracy compared to previous studies that did not apply feature selection or used different datasets.

In research conducted by (Tasnim & Habiba, 2021), the accuracy of sentiment classification using Random Forest, SVM, and XGBoost was 92.85%, 82%, and 83.5%, respectively. Meanwhile, research conducted by (D. J. Putri et al., 2023) resulted accuracy of XGBoost reaching 75% accuracy after applying Feature Selection Chi-Square. In this study, which applied Feature Selection Chi-Square specifically to sentiment analysis in the Indodax application, the following accuracy results were obtained SVM reaching 94%, Random Forest reaching 91%, and XGBoost reaching 92%. These results show that the implementation of the model in this study has a competitive level of accuracy and is in line with previous findings. However, there are small variations that can be caused by differences in dataset characteristics, preprocessing methods, and parameter tuning strategies.

Although Feature Selection Chi-Square has been shown to improve the performance of the model, some limitations in this study need to be noted. First, the dataset used in this study focuses on sentiment analysis on Indodax, so the findings obtained may not be generalizable to other cases. Secondly, although feature selection helps eliminate irrelevant and redundant features, it does not completely eliminate potential bias in model learning, especially in dealing with complex linguistic structures. Thirdly, the process of optimal feature selection and hyperparameter tuning can be quite a computational burden, which needs to be carefully considered in large-scale implementations.

This study confirms that the application of Feature Selection Chi-Square can improve the performance of SVM, Random Forest, and XGBoost models in sentiment analysis tasks. Compared to previous studies, the results obtained show that the approach used in this study is effective, especially in analyzing financial sentiment on Indodax. Future research can explore other feature selection methods, such as information gain or deep learning-based feature extraction, to further optimize sentiment classification models.

CONCLUSION

The comparison of Support Vector Machine (SVM), Random Forest, and XGBoost models with Feature Selection Chi-Square shows that SVM has the most optimal performance. The accuracy of SVM increased from 91% to 94% after the application of Feature Selection Chi-Square, which successfully reduced the number of features from 52,312 to 2,000 without a significant decrease in accuracy. This proves that Feature Selection Chi-Square is effective in removing irrelevant features without losing important information. In addition, SVM shows superiority in handling data that has been filtered from redundant information, improving the efficiency of feature selection and producing the best performance in sentiment classification of Indodax reviews.

REFERENCES

- Afriyani, S., Surono, S., & Solihin, I. M. (2024). Chi-Square Feature Selection with Pseudo-Labeling in Natural Language Processing. *JTAM (Jurnal Teori Dan Aplikasi Matematika)*, 8(3), 896. <https://doi.org/10.31764/jtam.v8i3.22751>
- Aida Sapitri, I., Fikry, M., Sains dan Teknologi, F., & Islam Negeri Sultan Syarif Kasim Riau, U. (2023). PENGKLASIFIKASIAN SENTIMEN ULASAN APLIKASI WHATSAPP PADA GOOGLE PLAY STORE MENGGUNAKAN SUPPORT VECTOR MACHINE. *Jurnal TEKINKOM*, 6(1). <https://doi.org/10.37600/tekinkom.v6i1.773>
- Bhoria, H., Dhankhar, A., & Solanki, K. (2023). INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY Chi-Square Feature Selection Technique for Student's performance prediction. / *Indian Journal of Science and Technology*, 16(38), 3250–3257. <https://doi.org/10.17485/IJST/v16i38.921>
- Chen Sami, D., Sugiharto, A., & Jie, F. (2024). CHI SQUARE FEATURE SELECTION FOR IMPROVING SENTIMENT ANALYSIS OF NEWS DATA PRIVACY TREATS. *Journal of Theoretical and Applied Information Technology*, 102(18). www.jatit.org



- Erkamim, M., Suswadi, S., Subarkah, M. Z., & Widarti, E. (2023). Komparasi Algoritme Random Forest dan XGBoosting dalam Klasifikasi Performa UMKM. *Jurnal Sistem Informasi Bisnis*, 13(2), 127–134. <https://doi.org/10.21456/vol13iss2pp127-134>
- Ernayanti, T., Mustafid, M., Rusgiyono, A., & Hakim, A. R. (2023). PENGGUNAAN SELEKSI FITUR CHI-SQUARE DAN ALGORITMA MULTINOMIAL NAÏVE BAYES UNTUK ANALISIS SENTIMEN PELANGGAN TOKOPEDIA. *Jurnal Gaussian*, 11(4), 562–571. <https://doi.org/10.14710/j.gauss.11.4.562-571>
- Mat, S. R. T., Razak, M. F. A., Kahar, M. N. M., Arif, J. M., & Firdaus, A. (2022). A Bayesian probability model for Android malware detection. *ICT Express*, 8(3), 424–431. <https://doi.org/10.1016/j.icte.2021.09.003>
- Nurtikasari, Y., Syariful Alam, & Teguh Iman Hermanto. (2022). Analisis Sentimen Opini Masyarakat Terhadap Film Pada Platform Twitter Menggunakan Algoritma Naive Bayes. *INSOLOGI: Jurnal Sains Dan Teknologi*, 1(4), 411–423. <https://doi.org/10.55123/insologi.v1i4.770>
- Oktavia, D., Ramadhan, Y. R., & Minarto. (2023). Analisis Sentimen Terhadap Penerapan Sistem E-Tilang Pada Media Sosial Twitter Menggunakan Algoritma Support Vector Machine (SVM). *KLIK: Kajian Ilmiah Informatika Dan Komputer*, 4(1), 407–417. <https://doi.org/10.30865/klik.v4i1.1040>
- Pasek, P., Mahawardana, O., Sasmita, G. A., Agus, P., & Pratama, E. (2022). Analisis Sentimen Berdasarkan Opini dari Media Sosial Twitter terhadap “Figure Pemimpin” Menggunakan Python. In *JITTER-Jurnal Ilmiah Teknologi dan Komputer* (Vol. 3, Issue 1).
- Permana, M. A., Widiastuti, S., & Saepudin, S. (2023). ANALISIS SENTIMEN PENGGUNAAN APLIKASI VIDEO CONFERENCE PADA ULASAN GOOGLE PLAY STORE MENGGUNAKAN METODE NBC (NAIVE BAYES CLASSIFIER). *JURSISTEKNI (Jurnal Sistem Informasi Dan Teknologi Informasi)*, 178–191.
- Petronella Purba, M., & Transver Wijaya, Y. (n.d.). *Analisis Basic Emotion Masyarakat Pada Masa Pandemi COVID-19 dengan Metode LSTM-FastText.....(Purba dan Wijaya) Analisis Basic Emotion Masyarakat Pada Masa Pandemi COVID-19 di Media Sosial Twitter Dengan Metode LSTM-FastText (Basic Emotion Analysis Of Community During The Covid-19 Pandemic On Twitter Social Media Using The LSTM-FastText Method)*.
- Putri, D. J., Dwifabri, M., & Adiwijaya, A. (2023). Text Classification of Indonesian Translated Hadith Using XGBoost Model and Chi-Square Feature Selection. *Building of Informatics, Technology and Science (BITS)*, 4(4). <https://doi.org/10.47065/bits.v4i4.2944>
- Putri, N. B., & Wijayanto, A. W. (2022). Analisis Komparasi Algoritma Klasifikasi Data Mining Dalam Klasifikasi Website Phishing. *Komputika : Jurnal Sistem Komputer*, 11(1), 59–66. <https://doi.org/10.34010/komputika.v11i1.4350>
- Repina Kesuma, P., Dewi, T., Kusumanto, R. D., Risma, P., Oktarina, Y., Elektro, J. T., Sriwijaya, N., & Id, T. C. (2020). *JOURNAL OF APPLIED SMART ELECTRICAL NETWORK AND SYSTEMS (JASENS) Desain Fuzzy Logic Controller Untuk Pengendali Pergerakan Mobile Manipulator* (Vol. 1, Issue 1). <http://journal.isas.or.id/index.php/JASENS>
- Rivanie, T., Pebrianto, R., Hidayat, T., Bayhaqy, A., Gata, W., & Novitasari, B. (2021). ANALISIS SENTIMEN TERHADAP KINERJA MENTERI KESEHATAN INDONESIA SELAMA PANDEMI COVID-19. In *Jurnal Informatika* (Vol. 21, Issue 1).
- Sartina Jayanti, T., Habibi, C., & Setiana, E. (n.d.). Analisis Sentimen Penggunaan Aplikasi Traveloka di Twitter Menggunakan Model Klasifikasi. 2024.
- Sunarko, G. S., Wasino, J., & Sutrisno, T. (2023). *KLASTERISASI SENTIMEN ULASAN PENGGUNA APLIKASI BCA MOBILE PADA PLATFORM GOOGLE PLAY STORE DENGAN ALGORITMA K-MEANS CLUSTERING*.
- Tasnim, F., & Habiba, S. U. (2021). A Comparative Study on Heart Disease Prediction Using Data Mining Techniques and Feature Selection. *International Conference on Robotics, Electrical and Signal Processing Techniques*, 338–341. <https://doi.org/10.1109/ICREST51555.2021.9331158>
- Wahyudi, R., Kusumawardhana, G., Purwokerto, A., Letjend, J., Soemarto, P., Purwanegara, K., Purwokerto, T., & Banyumas, K. (2021). Analisis Sentimen pada review Aplikasi Grab di Google Play Store Menggunakan Support Vector Machine. *JURNAL INFORMATIKA*, 8(2). <http://ejournal.bsi.ac.id/ejurnal/index.php/ji>
- Yasir, M., Grace Haque, M., Suraji, R., & Author, C. (n.d.). Analisis Sentimen Terhadap Kontroversi Fatwa MUI Nomor 83 Tahun 2023 Tentang Pemboikotan Produk yang Terafiliasi Israel. 2024. <https://doi.org/10.31933/jemsi.v5i4>
- Yudha, I. G. L., Sanjaya ER, N. A., Karyawati, A. A. I. N. E., Dwidasmara, I. B. G., Putra, I. G. N. A. C., & Mahendra, I. B. M. (2022). Article Classification Using Convolutional Neural Network (CNN) And Chi-Square Feature Selection. *JELIKU (Jurnal Elektronik Ilmu Komputer Udayana)*, 11(3), 529.