p–ISSN: 2723 - 6609 e-ISSN: 2745-5254
Vol. 5, No. 10, October 2024 http://jist.publikasiindonesia.id/

Enhancing XGBoost Classification with SVM-SMOTE & EasyEnsemble for Imbalanced Telemedicine Sentiment Data

Ahmad Yusran Siregar[1*], Ajib Setyo Arifin[2]
Universitas Indonesia, Indonesia
Email: ahmad.yusran01@ui.ac.id[1*], ajib.sa@ui.ac.id[2]

*Correspondence

Keywords: telemedicine, imbalance data, xgboost, svm-smote, easy ensemble.

ABSTRACT

Telemedicine is the practice of health through applications using audio, visual, and data communication, including care, diagnosis, consultation, and treatment as well as remote medical data exchange. Based on the results of sentiment analysis on telemedicine applications, imbalanced data is often found. The purpose of this research is to identify the use of SVM-SMOTE and EasyEnsemble in improving the performance of XGBoost classification on sentiment data imbalance in Telemedicine. Identification is done by including SVM-SMOTE and EasyEnsemble methods in improving XGBoost Classification Performance using data obtained from the Halodoc application, then validation techniques will be carried out using AUC and GMeans. The results showed that the use of SVM SMOTE and EasyEnsamble for data imbalance in XGBoost obtained the best model that is feasible to use in improving the performance of imbalance classification of sentiment data in health applications. Conclusion The test results show that the combination of Xgboost, SVM-SMOTE, and EasyEnsemble produces the best model to improve the performance of unbalanced sentiment data classification. It is recommended to add a neutral label in the sentiment analysis for comparison with this study.

Introduction

In Indonesia, the development of health mobile apps that provide telemedicine services began in 2015 and regulations governing telemedicine were made in 2019 (Jamil et al., 2015). The number of users of health-based digital applications amounted to 10% of the total population of Indonesia in 2019. Telemedicine is rapidly evolving as a cost-effective and efficient service to meet the increasing need for high-quality healthcare, especially during the COVID-19 pandemic. During the COVID-19 pandemic, the app also provided COVID-19 vaccination and testing programs to support the government in breaking the chain of transmission of the Coronavirus. Despite their advancements, telemedicine apps must enhance their quality and assess their performance through sentiment analysis of their products. Companies can utilize reviews from the Google Play

store to gauge public sentiment about their app versions. Sentiment analysis involves examining text to determine moods, emotions, or attitudes toward products, services, individuals, organizations, and events. The sentiment is categorized as positive, negative, or neutral. This analysis is frequently employed to refine decision-making and boost customer satisfaction within a company's business operations (Safitri et al., 2021). In previous research with the title Sentiment Analysis on Telemedicine App Reviews using XGBoost Classifier conducted by (Afifah et al., 2021) sentiment analysis was carried out on halodoc telemedicine using XGBoost and an accuracy rate of 96.24% was obtained. There are 4 factors of negative sentiment, namely payment, place, service, and system.

However, the study explained that there was imbalance or unbalanced data and that no technique was used to handle unbalanced data in the study. Therefore, it is necessary to develop conduct analysis using different modification algorithms to get better accuracy results than previous studies, especially in handling imbalanced data found in sentiment analysis. Data imbalance occurs when samples are distributed unevenly across different categories within a dataset. The dataset can be categorized into a majority class and a minority class based on the size of the samples (Zhang et al., 2022). Learning classifiers from imbalanced data sets is a common problem in classification problems, which is a serious problem. In this situation, the majority of examples belong to one class, while the other class, which usually consists of more important traits, actually accounts for a much lower number of examples (Tyagi & Mittal, 2020). XGBoost, a new and effective ensemble learning algorithm, is extensively used for its numerous benefits, though its performance in classifying imbalanced data is frequently suboptimal (Haixiang et al., 2017).

The extreme gradient boosting algorithm, XGBoost, is an ensemble learning method known for its high flexibility, strong predictive power, excellent generalization capabilities, scalability, efficient model training, and robustness (He et al., 2021). For research involving imbalanced data, the enhanced SMOTE algorithm can be integrated with XGBoost for clustering, utilizing ensemble learning to detect anomalies in the bolting process (Zhang et al., 2022).

In addition to carrying out the optimization process in improving the performance of the XGBoost classification in solving problems in data imbalance, this study also carried out the process of optimizing the management output of reviews obtained in telemedicine applications with the help of K-means (Henriques et al., 2020). Large volumes of data may be systematically accumulated across several data-collecting locations thanks to recent advancements in big data science data-gathering techniques (Vankayalapati et al., 2021). The most widely used and straightforward clustering algorithm is still K-means. This algorithm's low computing complexity and ease of implementation are the reasons for its widespread use in several clustering application fields (Ikotun et al., 2023). The process of correctly organizing unlabeled data is a part of data mining that is handled by cluster analysis utilizing K-means to generate data output in clusters to assist the process of generating suggestions from the resultant management (Capó et al., 2020).

Sentiment analysis in telemedicine has evolved along with the increasing use of digital health platforms. Several studies use classical methods such as Naive Bayes and Support Vector Machine (SVM) to classify patient sentiment towards digital-based health services. However, this study is still dealing with the problem of data imbalance, where positive sentiment is more dominant than negative or neutral sentiment.

Meanwhile, EasyEnsemble (Liu et al., 2009) introduced an ensemble technique to overcome data imbalance. This technique focuses more on the approach of combining multiple selectively trained models on a minority subset of data to improve classification capabilities.

Most previous studies used SMOTE directly without modification, while this study used SVM-SMOTE, which combines the advantages of SVM in establishing a more accurate minority class margin before the oversampling process, reducing the possibility of noise and overfitting.

The combination of SVM-SMOTE and EasyEnsemble in the XGBoost algorithm is a new approach that has not been widely applied in sentiment analysis in the field of telemedicine. Most previous sentiment analysis studies have only used one oversampling or ensemble approach, without combining the two methods.

While there has been research on sentiment analysis using XGBoost, this study focuses on telemedicine, a domain that has not been explored extensively in the context of unbalanced data classification. This study fills a literature gap on how to improve the accuracy of the model in telemedicine sentiment analysis with such a hybrid approach.

Thus, this study offers an innovative approach by combining several techniques in dealing with the unique data imbalance problem in the telemedicine sentiment analysis sector, which makes an important contribution to the development of more accurate and effective classification methods.

The main purpose of using this algorithm is to group unlabeled data so that data objects that have the same characteristics and attributes are in one cluster so that the similarity of data objects in the same cluster is higher when compared to data objects from other clusters.

Method

Sentiment analysis to collect data on several reviews of the Halodoc platform or application through Google Playstore to get positive and negative reviews. In collecting sentiment data, there are five stages carried out, the first stage is data collection. Following data collection, each review undergoes a sentiment labeling process. Subsequently, the dataset undergoes data preprocessing, which involves case folding, punctuation removal, tokenization, stopword removal, normalization, and stemming. This preprocessing aims to eliminate noise, making the text clearer and more comprehensible. Next, a feature extraction process converts the text data into numerical or vector data. Additionally, the dataset is partitioned into training and test sets in a 75:25 ratio.

Using Python's Google Play Scraper Library, data is gathered from the Google Play Store during the data gathering phase. 12,785 review data in total in 2021. Username,

content, rating score, date, and response are all included in the review data. Nevertheless, the content column and rating score are the only ones used in this study. The data gathering procedure is part of the Data Cleaning & Labelling step, and the data is stored in a.csv file type. Before the labeling procedure is initiated, reviews that contain punctuation or emojis that are not required for the analysis are removed. Once superfluous data has been removed, the rating score is mapped using the Google Play Store's definition to complete the tagging process.

Score 1 for negative, 2 for somewhat negative, 3 for neutral, 4 for moderately positive, and 5 for positive. In this study, only negative and positive labels were taken with a total of 10,306 data.

Table 1
Number of Reviews Based on Sentiment Data

No	Sentiment	Number of Reviews
1	Positive	9458
2	Negative	848

During the data preprocessing phase in language processing, the information typically involves unstructured data with significant noise. Thus, it's imperative to convert this data into a structured format before proceeding. For this research, the data comprises Indonesian-language reviews. Therefore, in this preprocessing stage, we utilize the Python Sastrawi library, available at github.com/har07/PySastrawi, a straightforward tool designed for tasks like stopword removal and stemming.

XGBoost Classifier

For regression and classification, the Xtreme Gradient Boosting technique, or XGBoost, is a useful approach. XGBoost leverages many residual rounds of value fitting to enhance machine learning performance and efficiency. It is built on the Gradient Boosting framework and adds a new Decision Tree regularly (Sagi & Rokach, 2021). The gradient-boosting decision tree technique is implemented by XGBoost. Boosting is an ensemble strategy in which flaws in previous models are corrected by adding new models. Up until no further advancements are possible, models are added one after the other.

SVM-SMOTE

The Synthetic Minority Oversampling Technique (SMOTE) is one of the methods used to handle imbalanced data problems. Synthetic Minority Oversampling Technique (SMOTE) uses minority data and creates synthetic data from it. The initial concept of the SMOTE algorithm is to identify the minority class and the majority class based on the number of classes in the data. After the minority class is identified (class 0), SMOTE will then calculate the synthetic data that will be formed based on the formula $Ty = (N/100) * Tx$. The variable Tx is the number of minority class data, Ty is the final minority class (after adding synthetic data), and N is the oversampling percentage with a multiple of 100 (Fonseca et al., 2021).

EasyEnsemble

An algorithm for hybrid ensemble undersampling is called Easyensemble. It compensates for the drawbacks of popular under-sampling techniques that could overlook crucial classification information by combining random sampling and the AdaBoost algorithm using Bagging. Additionally, the underlying classifier's choice of the AdaBoost method enhances both generalization and classification accuracy. EasyEnsemble is based on unsupervised sampling, in contrast to the supervised combination of another representative method, BalanceCascade. This method can prevent the wasting of scarce data resources because of its benefits in low time complexity and high data utilization (Qian et al., 2024).

K-Means

K-Means is a data mining algorithm commonly employed for data clustering. It operates on numeric attributes, utilizing a distance-based approach to partition data into clusters. The K-Means clustering algorithm is widely recognized and known for its flexibility, efficiency, and straightforward implementation (Cui, 2020).

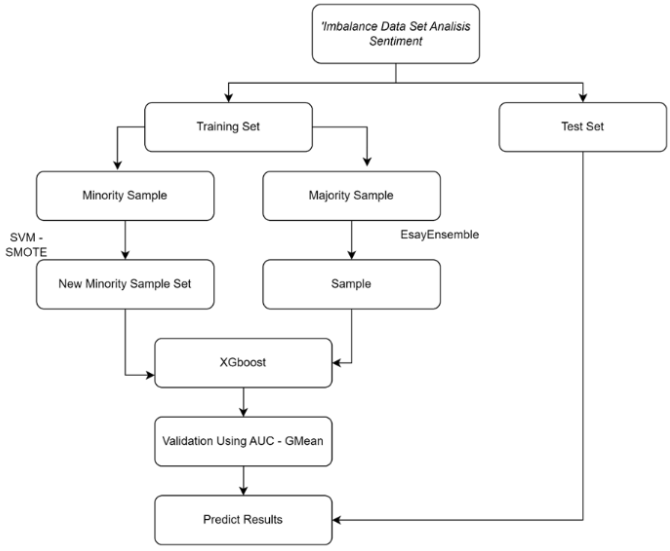

Fig. 1 Flowchart of Research

Results and Discussion

Model Performance Evaluation

The data used in this study is primary data, namely data on reviews or reviews obtained from the results of checking through the Google Play Store application on the Halodoc application from January to September 2023 with a total of 12784 reviews or reviews.

To overcome imbalanced data using Xgboost, SVM SMOTE, and EasyEnsamble. SMOTE is done only on training data. If SMOTE is performed on training data and testing data, the results obtained will be too good or over-optimistic because the testing data is replicated with the same data pattern in the training data and testing data. Model evaluation process using K-Fold Cross-Validation and confusion matrix. The results of

each K-fold cross-validation can be seen in Table 2. The classification results of the Xgboost, SVM SMOTE, and EasyEnsamble methods can be seen in Table 2

Table 2
Classification Results Of K-Fold Data

No	Replication	F1 Score
1	I	0.9633
2	II	0.9732
3	III	0.9749
4	IV	0.9748
5	V	0.9600
6	VI	0.9704
7	VII	0.9710
8	VIII	0.9638
9	IX	0.9756
10	X	0.9739

Table II shows the results of sentiment classification in the XGboost health application before the data imbalance technique is performed, using K-Fold validation from the results of evaluating the accuracy of the mode with the first to ten repetitions, an average accuracy value of 97% is obtained.

Table 3
Classification Results of Xgboost, Svm-Smote, and Easyensemble Method

XGBoost		Xgboost+SVM SMOTE,EasyEnsamble	
AUC	**GMeans**	**AUC**	**Gmeans**
0.8670	0.7958	0.8992	0.8156

Table 3 shows the comparison of accuracy values using AUC and GMeans on the XGboost Algorithm before and after data imbalance techniques using SVM-SMOTE and EasyEnsemble. From the results obtained, by only using the XGBoost Algorithm, the

AUC value is 86.7% and GMeans is 79.58%. After the data imbalance technique, it shows more accuracy with an AUC value of 89.92% and GMeans 81.56%.

Visualization Data and Clustering

Based on the data visualization and clustering process that has been processed using the K-Means algorithm, there are 4 groups as keywords used in each cluster obtained. The 4 keywords consist of Service with 674 words, Payment with 475 words, System with 671 words, and Place with 313 words. In this study, the authors used the fishbone diagram in Figure 2. to identify factors that must be done to improve or improve the quality of service of Health Applications (Telemedicine), especially Halodoc from the results of negative reviews and determine recommendations from the output obtained.

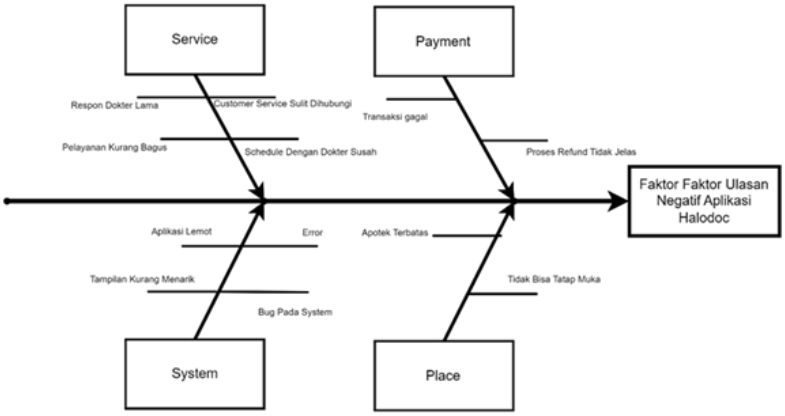

Fig 2 Fishbone Diagram for Negative Sentiment

The author found four main factors that users complained about the most which include the following aspects; payment, venue, service, and system. Users complained about difficult payment methods and refunds that took too long. Users also complained about services or pharmacies where drug pickup is not yet available in some areas, incomplete order items, and crashes or errors that sometimes occur on the application. Based on these findings, the authors provide recommendations to improve the quality of service provided to customers based on these 4 aspects. In the service aspect, it is necessary to improve services, especially those related to customer service so that if there are problems, users can be immediately given a solution. Furthermore, in the payment aspect, it is necessary to improve and evaluate the payment method used so that payment transactions can run smoothly. In the next aspect, namely the system, it is necessary to use a simple interface to make it easier for users to use the application and avoid loading data that causes errors. For the place aspect, it is necessary to develop a collaborative process with many pharmacies or health services in all regions so that affordability or application access can be evenly distributed.

Conclusion

Aspects used in this study include Xgboost, SVM SMOTE, and EasyEnsamble on imbalance data obtained from sentiment analysis in health or telemedicine applications, division of data schemes 75% training data and 25% testing data into the best scheme for data division, models with the best parameters. SVM test results using data that has been balanced with Xgboost and EasyEnsamble obtained the best model that is feasible to use in improving the classification performance of imbalanced sentiment data in health applications. Xgboost method provides excellent classification performance on imbalanced data through improved classification performance using SVM-SMOTE and EasyEnsemble. We suggest including a neutral label in sentiment analysis and comparing it to this study.

Bibliography

Afifah, K., Yulita, I. N., & Sarathan, I. (2021). Sentiment analysis on telemedicine app reviews using xgboost classifier. *2021 International Conference on Artificial Intelligence and Big Data Analytics*, 22–27.

Capó, M., Pérez, A., & Lozano, J. A. (2020). An efficient K-means clustering algorithm for tall data. *Data Mining and Knowledge Discovery*, *34*, 776–811.

Cui, M. (2020). Introduction to the k-means clustering algorithm based on the elbow method. *Accounting, Auditing and Finance*, *1*(1), 5–8.

Fonseca, J., Douzas, G., & Bacao, F. (2021). Improving imbalanced land cover classification with K-Means SMOTE: detecting and oversampling distinctive minority spectral signatures. *Information*, *12*(7), 266.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220–239.

He, S., Li, B., Peng, H., Xin, J., & Zhang, E. (2021). An effective cost-sensitive XGBoost method for malicious URLs detection in imbalanced dataset. *IEEE Access*, *9*, 93089–93096.

Henriques, J., Caldeira, F., Cruz, T., & Simões, P. (2020). Combining k-means and xgboost models for anomaly detection using log datasets. *Electronics*, *9*(7), 1164.

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178–210.

Jamil, M., Khairan, A., & Fuad, A. (2015). Implementasi aplikasi telemedicine berbasis jejaring sosial dengan pemanfaatan teknologi cloud computing. *JEPIN (Jurnal Edukasi Dan Penelitian Informatika)*, *1*(1).

Qian, Y., Yao, S., Wu, T., Huang, Y., & Zeng, L. (2024). Improved Selective Deep-Learning-Based Clustering Ensemble. *Applied Sciences*, *14*(2), 719.

Safitri, R., Alfira, N., Tamitiadini, D., Dewi, W. W. A., & Febriani, N. (2021). *Analisis Sentimen: Metode Alternatif Penelitian Big Data*. Universitas Brawijaya Press.

Sagi, O., & Rokach, L. (2021). Approximating XGBoost with an interpretable decision tree. *Information Sciences*, *572*, 522–542.

Tyagi, S., & Mittal, S. (2020). Sampling approaches for imbalanced data classification problem in machine learning. *Proceedings of ICRIC 2019: Recent Innovations in Computing*, 209–221.

Vankayalapati, R., Ghutugade, K. B., Vannapuram, R., & Prasanna, B. P. S. (2021). K-Means algorithm for clustering of learners performance levels using machine learning techniques. *Rev. d'Intelligence Artif.*, *35*(1), 99–104.

Zhang, P., Jia, Y., & Shang, Y. (2022). Research and application of XGBoost in imbalanced data. *International Journal of Distributed Sensor Networks*, *18*(6), 15501329221106936.