Sentiment Analysis of Bali Calendar Application Reviews using K-Nearest Neighbour

Ni Made Dina Aprilianti*[,a,1], Rosalia Arum Kumalasanti[a,2]
[a]Sanata Dharma University, Yogyakarta, Indonesia

Corresponding Author: madedinaapril@gmail.com

Abstract

This study evaluates user sentiment towards the Bali Calendar application, analysing both positive feedback and negative critiques. The research employs the K-Nearest Neighbour (KNN) algorithm to classify sentiments as either positive or negative, aiming to assess overall public satisfaction with the app. To improve classification performance, the Tomek Links technique is applied in conjunction with KNN. The study categorizes data into pre- and post-COVID periods to address the observed increase in negative reviews following app updates during the pandemic. In the pre-COVID phase, KNN achieved accuracy rates of 93.7% and 94.3% with and without Tomek Links, respectively, using parameter values K=5 and K=3. In the post-COVID period, accuracy rates were 86.0% and 87.2% with and without Tomek Links, respectively, using parameter K=9. The application of Tomek Links resulted in a notable accuracy improvement of 1.2% in the post-COVID data compared to a 0.6% increase in the pre-COVID data. This finding highlights the significant role of Tomek Links in enhancing KNN accuracy, particularly when dealing with unbalanced datasets. The study demonstrates that while KNN performs robustly, Tomek Links can provide a substantial boost in classification accuracy, especially in scenarios with skewed data distributions.

Keywords: Balinese Calendar, Sentiment Analysis, K-Nearest Neighbour, Imbalance Dataset, Tomek Links

I. Introduction

Current technological developments have significantly changed the way humans meet their needs, bringing convenience, speed and practicality in access to various information and products. This development also influenced the way Balinese, especially those outside the island of Bali, gained access to the Balinese calendar, an important aspect of carrying out their traditions and culture. Before technological advances, getting a Balinese calendar for Balinese people outside the island of Bali was quite a difficult task. However, with the advent of information technology, especially the increase in internet use in Indonesia, access to the Balinese calendar has become easier, get the Balinese calendar through the Balinese Calendar application which is available on the Google Play Store.

The Bali Calendar application is a modern solution for Balinese people who live outside the island of Bali. This app not only simplifies the process of getting a Balinese calendar, but also allows users to stay connected with Balinese traditions and culture. This application provides information regarding *peting* dates, holidays and other celebrations. Features such as searching for *wariga* elements, matchmaking, and setting reminders make this application a very useful tool [2]. However, like many other applications, the Bali Calendar application also receives various responses from its users, both positive such as praise and negative in the form of user complaints. One platform that is a place for users to provide reviews is Google Play Store, reviews in the form of ratings and opinions regarding the Bali Calendar application.

Sentiment analysis of user reviews on the Google Play Store can provide deep insight into user evaluations and preferences which can improve the quality and user satisfaction with this application. In data processing for sentiment analysis, unbalanced data conditions are sometimes found, where the amount of data tends to be greater in certain classes. Unbalanced data will certainly cause problems in the sentiment analysis process because it can cause the sentiment analysis model to be less accurate. To overcome this problem, implementing a data balancing method such as Tomek Link can be a solution. Based on research conducted by Ida Ayu Mirah Cahya Dewi, I Komang Dharmandra and Ni Wayan Setiasih (2023) conducted a sentiment analysis of reviews of the Satu Sehat.

Mobile application using the KNN and Tomek Links methods. The results obtained by KNN classification without Tomek Links produced an accuracy of 74.91%. This result was less than optimal when compared to classification using KNN and Tomek Links which obtained an accuracy of 76.40% [9]. The classification method used for sentiment analysis is also very influential in getting optimal results. Based on research conducted by Puji Astuti and Nuzuliarini Nuris (2022) on reviews of the Peduli Protect application from the Google Play Store using the KNN method, it resulted in an accuracy of 81.74% [4].

Drawing from existing studies, this research seeks to classify sentiments in reviews of the Bali Calendar application on the Google Play Store using the K-Nearest Neighbour (KNN) method, incorporating Tomek Links to address data imbalances. The primary objective is to gain a comprehensive understanding of public perceptions and evaluations of the application. By employing this methodology, the research seeks to address the challenges associated with sentiment classification in the presence of skewed data distributions, thereby providing a clearer picture of user satisfaction and areas for improvement. Furthermore, this research endeavours to bridge the gap between information technology and local cultural knowledge. By leveraging technological advancements and incorporating traditional wisdom, the study aims to create innovative solutions for application development and evaluation. This integration is expected to yield new insights and contribute to the enhancement of digital culture, offering valuable perspectives on how applications can be developed and refined to better meet user needs and preferences. Ultimately, the research strives to add significant value to both the field of digital technology and the broader context of cultural engagement.

II. Method

A. Research Overview

The research began with collecting review data through the Google Play Store, using web scraping techniques with the `google-play-scraper` package in the Python programming language. Next, the data undergoes preprocessing, including labelling, case folding, cleaning, tokenizing, normalization, stemming, and stop word removal. The goal of pre-processing is to make review data ready for use in the next stage. After pre-processing, word weighting is carried out using the TF-IDF technique, which produces vectors between documents and words. The data is then divided into training data and testing data using k-Fold Cross Validation. The next stage involves the classification process using the KNN method, as well as classification using KNN with the application of Tomek Links to handle data imbalance. Evaluation is carried out through a confusion matrix, which includes accuracy, precision and recall, to measure the performance of the classification method used.

B. Data Collection

This research will use review data for the Bali Calendar application from the Google Play Store. Data collection was carried out using web scrapping techniques using the google-play-scraper package from the Python programming language. The review data that will be taken will be a review that uses Indonesian. Review data was obtained in Indonesian from August 4, 2013, to June 28 2023, with a total of 5617 reviews.

C. Pre-Processing

At this stage, the process of converting raw reviews into review data that is clean and ready to use is carried out. First, the data will be labelled based on the review score value, with a score of more than 3 it will be labelled positive sentiment, while a score of less than 3 will be labelled negative [20]. After the labelling process, the data will be converted to lowercase and remove characters other than letters that are not needed such as punctuation, numbers and symbols that are not relevant. Then the data will undergo a tokenizing process, in this process cutting words or separating sentences into words separated by commas is carried out. After going through this process, data that has redundant words or abbreviated words will be replaced with standard words in accordance with the rules in the Big Indonesian Dictionary (KBBI) [8]. After that, the affixes in the words in the dataset will be removed to become base words. This process is assisted by the Indonesian language stemmer from the literary package. The final step in the pre-processing process is removing conjunctions that have no meaning. The deletion of this word is based on the use of the Indonesian Stop word List dictionary.

D. Division of Data Time Range

The process of dividing the data time range is carried out by reading the pre-processing results, then separating the data into two periods: before and after the COVID-19 pandemic period. The pre-COVID time span includes reviews from August 1, 2013, to February 29, 2020, while the post-COVID time span includes reviews from March 2, 2020, to June 28, 2023. The purpose of this division is to explore indications of an increase in negative reviews following app updates during the pandemic.

E. Word Weighting

Word weighting is the process of giving weight to each word, where in this research, the technique used for word weighting is Term Frequency - Inverse Document Frequency (TF-IDF). The data is first read and then the TF value is calculated with the aim of finding terms or words that match the document and the DF value aims to find the number of documents containing the term, then calculates the IDF weight with the aim of knowing that the term being searched matches the keyword, then finally calculate TF-IDF weights to find out important or frequently appearing words in documents [15]. The TF-IDF results will later be used for the classification process.

F. Data Sharing

Before entering the classification process, the data is first divided into training data and testing data using k-Fold Cross Validation. The k values that will be used in this research are 3, 5, 7, and 9. The random state value used is 42. From the results of dividing the data with different k values, different amounts of training data and testing data will be obtained according to the division. the folds. This allows for accuracy values for different classifications, depending on the number of folds for dividing the data in the k-Fold Cross Validation process. To visualize the distribution of data before and after the Covid pandemic using different k values, you can see in Figure 1 and Figure 2.

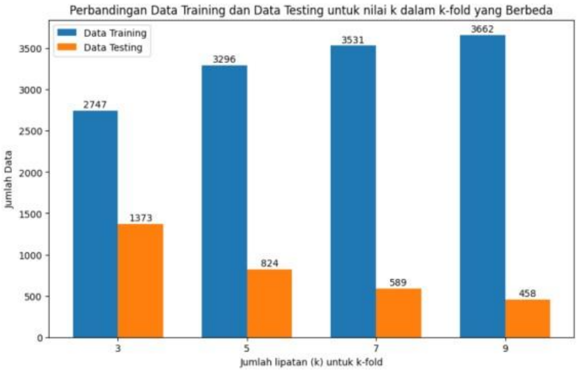


Figure 1 Distribution of Data Sharing before the Covid Pandemic

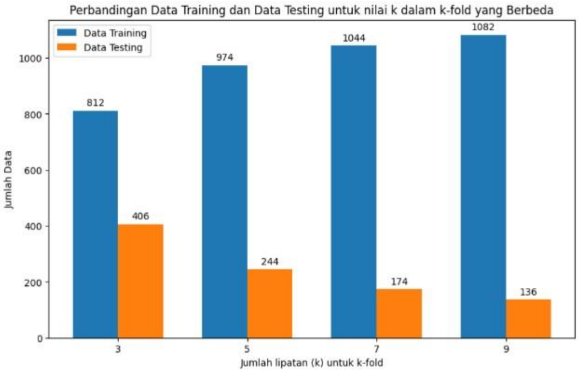


Figure 2 Distribution of data sharing after the Covid pandemic

G. Classification

The classification stage uses the KNN method, where KNN is a classification method that groups new data based on its distance to several data or nearest neighbours. The K/neighbour values used are 3, 5, 7, and 9. Classification is carried out for each data resulting from the division of k-Fold Cross Validation, so that classification is carried out four times for k-fold 3, 5, 7, and 9. The classification process is by KNN begins by initializing training data and test data resulting from data sharing. Next, input the K or neighbour values (3, 5, 7, and 9) and calculate the Euclidean distance between the training data and test data. After getting the distance, the training data is sorted based on smallest to largest distance. Then, K training data with the smallest distance from the test data are selected as the nearest neighbours. The majority label from the nearest neighbours will be used as a prediction for the test data. This process is repeated for each test data at each k-fold iteration.

H. Data Balancing

Data balancing is done with Tomek Links under sampling. Tomek Links balances the training data by reducing the number of majority samples whose positions are close to the minority samples around them. By using this balancing method, it is hoped that the comparison of the number of samples between the majority and minority classes will be more balanced, so that it can improve the quality of model learning, especially in the minority class. Comparison of data distribution before and after using Tomek Links on data before and after the Covid pandemic can be seen in the following table 1:

Table 1 Comparison of Data Distribution of Tomek Links before the Covid Pandemic

K-Fold	**With *Tomek Links***		**Without *Tomek Links***	
	Positive	**Negative**	**positive**	**Labelled**
3	2525	203	2544	203
5	3026	244	3052	244
7	3245	262	3270	262
9	3369	271	3392	271

Table 2 Comparison of Data Distribution of Tomek Links after the Covid Pandemic

K-Fold	With *Tomek Links*		Without *Tomek Links*	
	Positive	**Negative**	**positive**	**Labelled**
3	678	122	690	122
5	816	147	828	147
7	874	157	887	157
9	908	163	920	163

Based on the data from the table above for data before the Covid pandemic Tomek Links on k-fold data 3, 5, 7 and 9, respectively reduced the majority data by 0.7%, 0.8%, 0.7% and 0.6%. The largest percentage decrease in data due to using Tomek Links was in k-fold 5 data, namely 0.8%. Then for data after the Covid pandemic, Tomek Links on k-fold data 3, 5, 7 and 9, respectively reduced the majority data by 1.5%, 1.2%, 1.2% and 1.1%. The largest percentage decrease in data due to using Tomek Links was in k-fold 3 data, namely 1.5%. When compared with the decline in data before the Covid pandemic, Tomek Links data after the Covid pandemic decreased most data by a higher amount.

I. Classification Using Balanced Data

Following the initial classification conducted on an unbalanced dataset, the subsequent step involves performing classification on a dataset that has been balanced. This process retains the use of the K-Nearest Neighbour (KNN) algorithm as the primary classification method. However, the dataset for this step will include only the positively labelled data that has been refined through data balancing techniques using Tomek Links. By applying Tomek Links to address data imbalances, the dataset is adjusted to better represent positive labels, ensuring a more equitable distribution of classes. This balanced dataset will then be used in the KNN classification process to improve accuracy and reliability, allowing for a more precise assessment of sentiment and user feedback.

III. Results and Discussion

To find out the performance results of the model built in this research, we will use a confusion matrix, to find values for accuracy, precision, recall and F1-score. The aim is to find out the comparison of classification results for data before and after the Covid pandemic and to find out the comparison of the use of the Tomek Links balancing method on unbalanced data. From the tests that have been carried out, the highest accuracy was obtained using the KNN and Tomek Links methods, producing the highest accuracy of 94.3% for data before the Covid pandemic, using a value of k = 3 for KNN and a value of k = 9 for k-Fold Cross Validation. The results of all experiments for data before the Covid pandemic can be seen in Figures 3 to Figure 10.

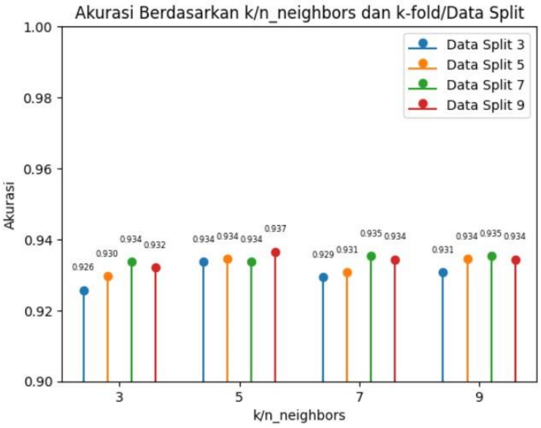


Figure 3 Comparison of Accuracy on Data before the Covid Pandemic

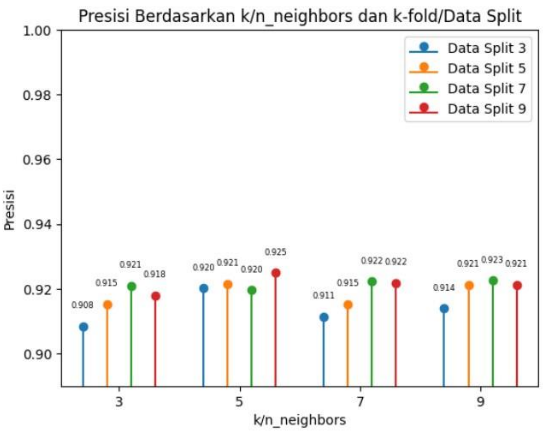
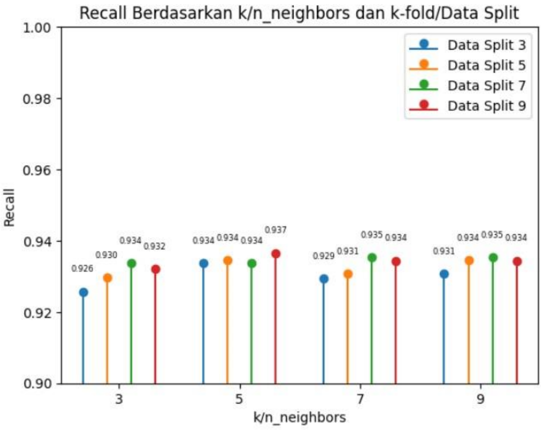


Figure 4 Comparison of Precision on Data before the Covid Pandemic



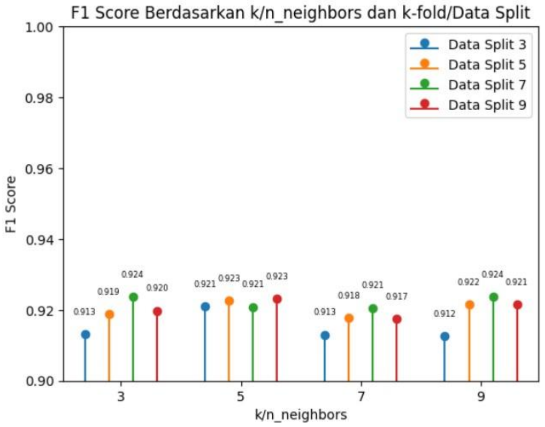


Figure 5 Comparison of Recalls on Data before the Covid Pandemic

Figure 6 Comparison of F1 Scores on Data before the Covid Pandemic

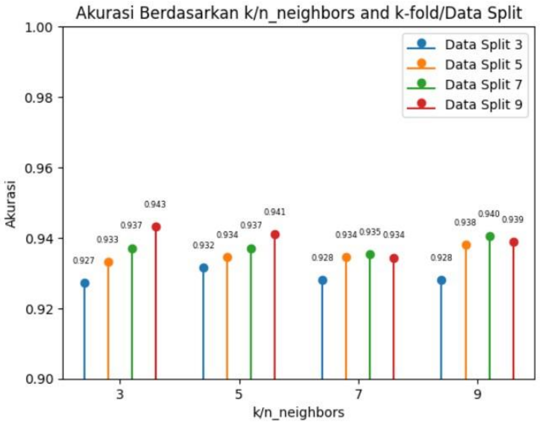


Figure 7 Comparison of KNN and Tomek Links Accuracy on Data before the Covid Pandemic

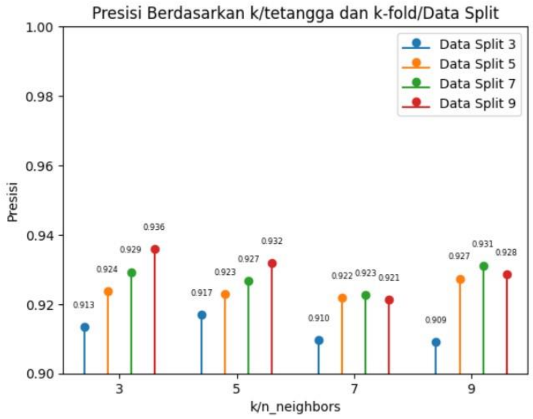


Figure 8 Comparison of KNN and Tomek Links Precision on Data before the Covid Pandemic

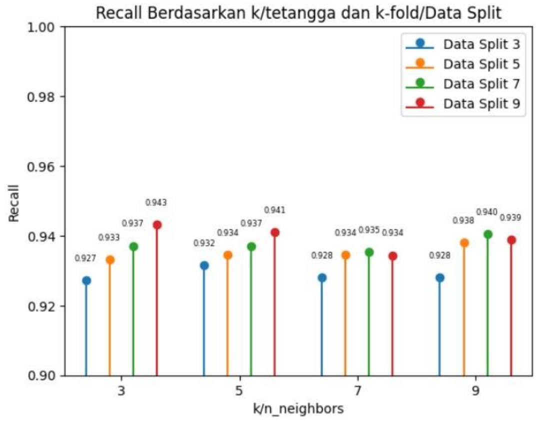


Figure 9 Comparison of KNN and Tomek Links Recalls on Data Before the Covid Pandemic

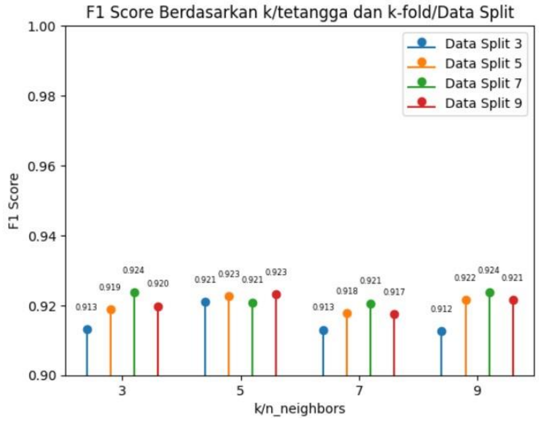


Figure 10 Comparison of KNN and Tomek Links F1 Scores on Data Before the Covid Pandemic

For data after the Covid pandemic, the highest accuracy was obtained by the KNN and Tomek Links methods which produced an accuracy of 87.2%, using a value of k = 9 for KNN and a value of k = 3 for k-Fold Cross Validation. The results of all experiments for data after the Covid pandemic can be seen in Figure 11 to Figure18.

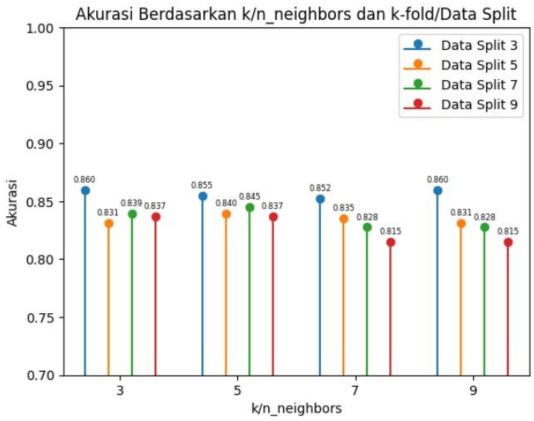


Figure 11 Comparison of Accuracy on Data After the Covid Pandemic

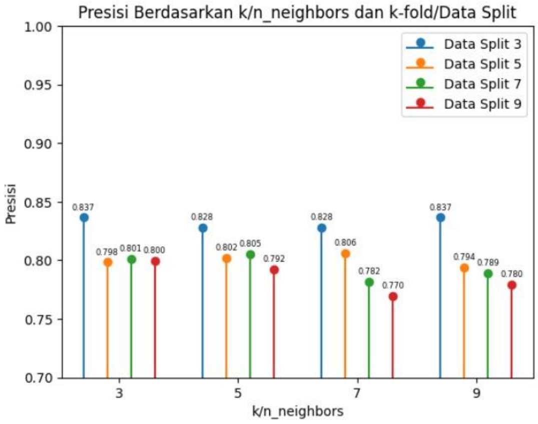


Figure 12 Comparison of Precision on Data Before the Covid Pandemic

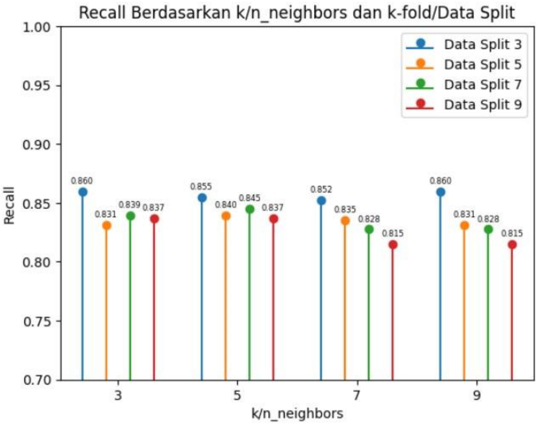


Figure 13 Comparison of Recalls on Data after the Covid Pandemic

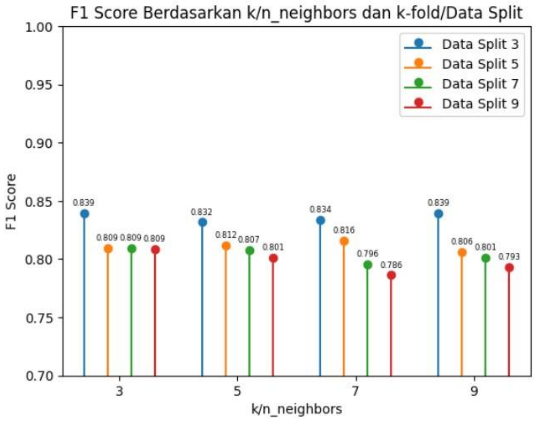


Figure 14 Comparison of F1 Scores on Data after the Covid Pandemic

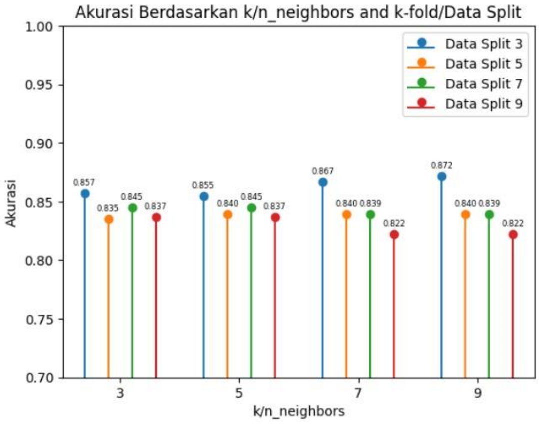


Figure 15 Comparison of KNN and Tomek Links Accuracy on Data after the Covid Pandemic

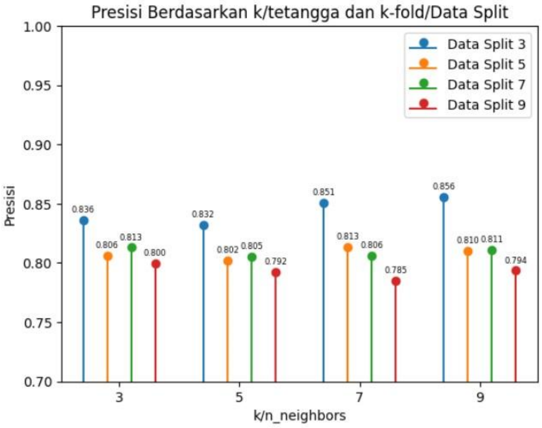


Figure 16 Comparison of KNN and Tomek Links Precision on Data after the Covid Pandemic

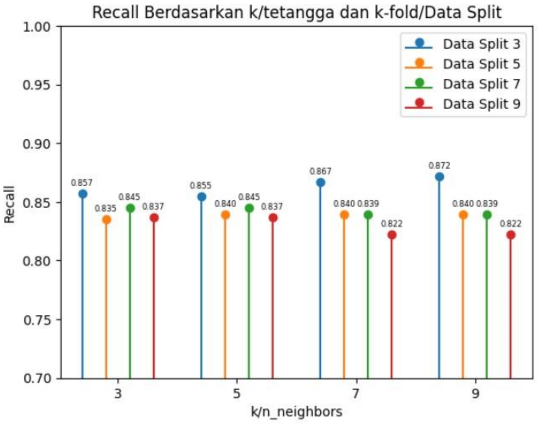


Figure 17 Comparison of KNN and Tomek Links Recalls on Data after the Covid Pandemic

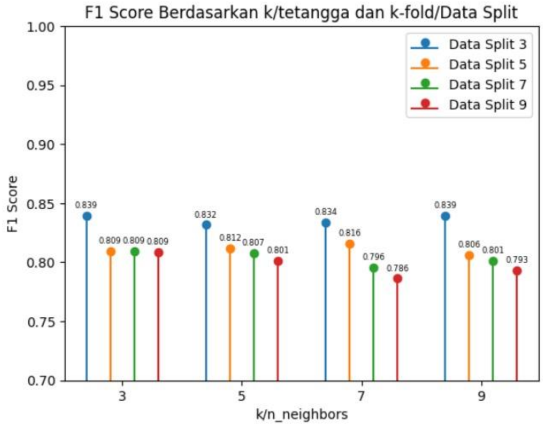


Figure 18 Comparison of KNN and Tomek Links F1 Scores on Data after the Covid Pandemic

The study explored the effectiveness of the Tomek Links technique in mitigating data imbalance, a common issue that can distort the outcomes of machine learning models by allowing the majority class to dominate. Prior to the COVID-19 pandemic, the application of Tomek Links resulted in a modest reduction in the proportion of majority class data, ranging between 0.6% to 0.8% depending on the k-Fold value used. The highest reduction, 0.8%, was observed at a k-Fold value of 5. Following the onset of the pandemic, however, the method showed a more pronounced impact. The data demonstrated a significant reduction in the majority class by up to 1.5% at a k-Fold value of 3, with other k-Fold values reflecting a reduction between 1.1% and 1.2%. This increased reduction suggests that the Tomek Links method became more effective in handling data imbalances during the post-pandemic period, likely due to the greater diversity or complexity of the data collected in that time frame.

To enhance the dataset further, the study employed the K-Nearest Neighbour (KNN) algorithm on both the original unbalanced data and a balanced version achieved using Tomek Links. The balancing process provided a more equitable representation of the minority class (positive labels), thereby improving the precision of sentiment analysis conducted on the dataset. The research also provided a comparative analysis of classification performance before and after the pandemic. Before the pandemic, the KNN algorithm alone reached its peak accuracy of 93.7% when k=5 for KNN and k=9 for k-Fold Cross Validation. With the introduction of Tomek Links, the accuracy rose to 94.3%, using k=3 for KNN and k=9 for k-Fold Cross Validation.

After the pandemic, the highest accuracy achieved by the KNN algorithm alone was 86.0%, with k=9 for KNN and k=3 for k-Fold Cross Validation. When combined with the Tomek Links method, this accuracy improved to 87.2%, maintaining the same parameter settings. In short, the combination of KNN and Tomek Links enhanced accuracy by 0.6% before the pandemic and by 1.2% after. These findings suggest that the Tomek Links technique was particularly advantageous during periods characterized by more complex or varied data distributions, such as those encountered after the pandemic.

IV. Conclusion

The study concludes that integrating the Tomek Links technique with the KNN algorithm can significantly improve the accuracy of sentiment classification tasks. The positive impact of Tomek Links was more evident in the post-pandemic period, highlighting its value in managing datasets that are more variable or imbalanced. This demonstrates the potential of Tomek Links to enhance the reliability of machine learning models in environments that are subject to frequent changes. The findings from this study emphasize the critical role of data balancing techniques like Tomek Links in real-world applications, where datasets often suffer from imbalances that can bias results. By comparing the data from before and after the COVID-19 pandemic, the study illustrates how external factors can alter data characteristics, necessitating robust techniques to manage these changes effectively. The increase in classification accuracy achieved by applying Tomek Links shows that such balancing methods can not only refine the dataset for better representation but also improve the predictive performance of machine learning models. This is particularly important in applications involving sentiment analysis or user feedback, where reliable outcomes are essential for informed decision-making in fields such as marketing, customer service, and public relations. In short, this current study reinforces the need for adaptive strategies in data management and suggests further research into other balancing techniques that could complement or enhance the effectiveness of Tomek Links, especially in scenarios characterized by increased complexity, such as those observed in the post-pandemic period.

REFERENCES

[1] Alhaqq, R.I., Putra, I.M.K. and Ruldeviyani, Y. (2022) ‘Analisis Sentimen terhadap Penggunaan Aplikasi MySAPK BKN di Google Play Store’, *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, 11(2), pp. 105–113. Available at: https://doi.org/10.22146/jnteti.v11i2.3528.

[2] Alitmd (2021) *Kalender Bali*. Available at: https://play.google.com/store/apps/details?id=com.alitmd.kalenderbali&hl=id&gl=US (Accessed: 23 May 2023).

[3] Aprilianti, N.M.D. *et al.* (2023) ‘Analisis Perbandingan Algoritma KNN, Gaussian Naive Bayes, Random Forest Untuk Data Tidak Seimbang Dan Data Yang Diseimbangkan Dengan Metode Tomek Link Undersampling Pada Dataset LCMS Tanaman Keladi Tikus’, 13(1), pp. 156–160.

[4] Astuti, P. and Nuris, N. (2022) ‘Penerapan Algoritma KNN Pada Analisis Sentimen Review Aplikasi Peduli Lindungi’, *Computer Science (CO-SCIENCE)*, 2(2), pp. 137–142. Available at: https://doi.org/10.31294/coscience.v2i2.1258.

[5] Badan Pusat Statistik (2019) *Proporsi Individu Yang Menggunakan Internet Menurut Provinsi*. Available at: https://www.bps.go.id/indicator/27/1225/1/proporsi-individu-yang-menggunakan-internet-menurutprovinsi.html (Accessed: 28 November 2023).

[6] Barus, S.G. (2022) ‘Klasifikasi Sentimen Data Tidak Seimbang Menggunakan Algoritma Smote Dan KNearest Neighbor Pada Ulasan Pengguna Aplikasi Pedulilindungi’, *Senamika*, pp. 162–173.

[7] Cholil, S.R. *et al.* (2021) ‘IJCIT (Indonesian Journal on Computer and Information Technology) Implementasi Algoritma Klasifikasi K-Nearest Neighbor (KNN) Untuk Klasifikasi Seleksi Penerima Beasiswa’, *IJCIT (Indonesian Journal on Computer and Information Technology)*, 6(2), pp. 118–127.

[8] Denes, I.M. *et al.* (1997) *Kamus Bahasa Indonesia-Bali A-K*. Jakarta: Pusat Pembinaan dan Pengembangan Bahasa.

[9] Dewi, I.A.M.C., Dharmendra, I.K. and Setiasih, N.W. (2023) ‘Analisis Sentimen Review Aplikasi Satu Sehat Mobile Menggunakan Model Sampling Tomek Links’, pp. 2–8. Available at: https://jurnal.undhirabali.ac.id/index.php/jutik/article/view/2644/3309.

[10] Erawan, L. (2015) ‘Pedoman Praktikum Standar Web’, *Pemrograman Web*, pp. 1–109.

[11] Iwandini, I., Triayudi, A. and Soepriyono, G. (2023) ‘Analisa Sentimen Pengguna Transportasi Jakarta Terhadap Transjakarta Menggunakan Metode Naives Bayes dan K-Nearest Neighbor’, *Journal of Information System Research (JOSH)*, 4(2), pp. 543–550. Available at: https://doi.org/10.47065/josh.v4i2.2937.

[12] Liu, B. (2015) *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Available at: https://doi.org/10.1017/CBO9781139084789.

[13] Mitchell, R. (2018) *Ryan Mitchell Web Scraping with Python*. Available at: www.allitebooks.com.

[14] Mondal, S. *et al.* (2023) ‘Machine Learning-based maternal health risk prediction model for IoMT framework’, *International Journal of Experimental Research and Review*, 32, pp. 145–159. Available at: https://doi.org/10.52756/ijerr.2023.v32.012.

[15] Purbo, O.W. (2019) *Text Mining, Analisis Medsos, Kekuatan Brand dan Intelijen di Internet*. ANDI.

[16] Python (2016) *Sastrawi*, *Python*. Available at: https://pypi.org/project/Sastrawi/ (Accessed: 27 November 2023).

[17] Python (2023) *Google-Play-Scraper*, *Python*. Available at: https://pypi.org/project/google-play-scraper/ (Accessed: 27 November 2023).

[18] Supono, R.A. and Suprayogi, M.A. (2021) ‘Perbandingan Metode TF-ABS dan TF-IDF Pada Klasifikasi Teks Helpdesk Menggunakan K-Nearest Neighbor’, *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 5(5), pp. 911–918. Available at: https://doi.org/10.29207/resti.v5i5.3403.

[19] Utami, H. (2022) ‘Analisis Sentimen dari Aplikasi Shopee Indonesia Menggunakan Metode Recurrent Neural Network’, *Indonesian Journal of Applied Statistics*, 5(1), p. 31. Available at: https://doi.org/10.13057/ijas.v5i1.56825.

[20] Veluchamy, A. *et al.* (2018) ‘Comparative Study cobagantiii of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches’, *SMU Data Science Review*, 1(4), pp. 1–22. Available at: https://scholar.smu.edu/cgi/viewcontent.cgi?article=1051&context=datasciencereview.