JPPIPA 10(8) (2024)

Jurnal Penelitian Pendidikan IPA

Journal of Research in Science Education

http://jppipa.unram.ac.id/index.php/jppipa/index

A Novel Hybrid Classification on Urban Opinion Using ROS-RF: A Machine Learning Approach

Usman Ependi[1*], Nahdatul Akma Ahmad[2]

[1] Informatics Department, Faculty of Science Technology, Palembang, Indonesia.
[2] College of Computing, Informatics, and Mathematics, Universiti Teknologi MARA, Malaysia.

Received: June 11, 2024
Revised: August 15, 2024
Accepted: August 25, 2024
Published: August 31, 2024

Corresponding Author:
Usman Ependi
u.ependi@binadarma.ac.id

DOI: 10.29303/jppipa.v10i8.8042

Abstract: Urban opinion from crowdsourced data often leads to imbalanced datasets due to the diversity of issues related to urban social, economic, and environmental topics. This study presents a novel hybrid approach that combines Random Over-Sampling and Random Forest (ROS-RF) to effectively classify such imbalanced data. Using crowdsourced urban opinion data from Jakarta, experimental results show that the ROS-RF method outperforms other approaches. The ROS-RF classifier achieved an impressive F1-score, recall, precision, and accuracy of 98%. These findings highlight the superior effectiveness of the ROS-RF method in classifying urban opinions, especially those related to social, economic, and environmental issues in urban settings. This hybrid approach provides a robust solution for managing imbalanced datasets, ensuring more accurate and reliable classification outcomes. The study underscores the potential of ROS-RF in enhancing urban data analysis and decision-making processes.

Keywords: Covid-19; Machine learning; Lexicon; Sentiment analysis

Introduction

In the constantly evolving realm of urban environments, a profound understanding of these settings is increasingly pivotal, especially in the quest to ameliorate the living conditions of their inhabitants. This imperative is accentuated considering that urban dwellers, who make up over half of the world's population, account for approximately 75% of global energy consumption (Kourtzanidis et al., 2021). In this dynamic context, city administrations are not merely reactive entities but proactive catalysts, spearheading a myriad of innovative endeavors through policy reform and the integration of cutting-edge information technology. These initiatives span from the visionary concept of smart cities to the practical implementation of sophisticated electronic governance systems (Ependi, Rochim, and Wibowo 2023b).

Equally critical to these innovative strides of city authorities is their alignment with the insights and recommendations emanating from the urban populace. The engagement of city residents is not just advantageous but effective for a understanding of urban requirements. Their discerning perspectives on varied urban dimensions, encompassing the societal, economic, and environmental spheres, are fundamental in fostering a robust trust in the policymaking process (Tomor et al., 2019). This strategy is in harmony with the quintessential principles of urban governance, where a synergetic interplay among the residents, city officials, and technological advancements forms the very foundation of urban development (Tomor et al. 2019; Castelnovo et al., 2016). Therefore, the active and meaningful participation of citizens is not just desirable but imperative. Their contributions and viewpoints serve as pivotal drivers in steering urban initiatives towards tangible success and palpable improvements in quality of life (Webster & Leleux, 2018).

The active involvement of citizens in the sphere of urban innovation emerges as a critical component for ensuring the sustainability and resilience of urban landscapes (Tan & Taeihagh 2020). Such engagement

How to Cite:
Ependi, U., & Ahmad, N. A. (2024). A Novel Hybrid Classification on Urban Opinion Using ROS-RF: A Machine Learning Approach. *Jurnal Penelitian Pendidikan IPA*, *10*(8), 5816–5824. https://doi.org/10.29303/jppipa.v10i8.8042

transcends beyond facilitating the creation of efficient urban systems; it cultivates a deeper sense of collective ownership and a shared responsibility towards the urban environment. When engaged, citizens are transformed from mere beneficiaries to key contributors and informants. Their feedback on the efficacy of urban programs Viale Pereira et al. (2017), Joia and Kuhl (2019), integral in ensuring these initiatives resonate with their actual needs, manifests in the form of crowdsourced data. This data, a rich tapestry of bottom-up information contributions extending beyond geographical confines Crooks et al. (2015), has demonstrated its prowess in shedding light on complex urban activities and addressing multifaceted challenges, areas where traditional datasets may falter (Long & Liu 2016). This repository of insights forms a linchpin in decision-making, propelling a more collaborative and inclusive approach towards crafting urban spaces that are not only better and more sustainable but also reflective of the aspirations and needs of their inhabitants.

Recognizing the indispensable role of crowdsourced data in empowering city authorities in their policymaking endeavors and in the advancement of information technology innovations necessitates a nuanced exploration. A paramount challenge in this domain is the discernment of the sentiment embedded in residents' feedback, positive or negative, which is critical for precise and effective decision-making. Accurate classification, therefore, becomes a cornerstone in this process. Pioneering studies in urban opinion classification, particularly in areas such as green spaces and flood management, offer illuminating insights. In the domain of green spaces, classification aids in deciphering residents' perceptions of the quality of urban greenery Ghahramani et al. (2021), while in flood management, it evaluates views on urban flood policies and response strategies (Wang et al., 2024).

Furthermore, the tapestry of urban life is intricately woven with threads of sustainability, encompassing social, economic, and environmental strands. This study, therefore, embarks on a mission to classify urban opinions employing sophisticated methodologies like random oversampler (ROS) and random forest (RF). These approaches are strategically selected to unravel and comprehend the nuanced perspectives of residents on the multifaceted aspects of urban social, economic, and environmental. ROS is distinguished for its proficiency in managing multi-class data (Ependi, Rochim, and Wibowo 2023a), while random forest is lauded for its effectiveness in text classification (Salles et al., 2018; Jalal et al., 2022; Chen et al., 2022). Through this study, we aim to contribute to a more nuanced understanding of urban dynamics, thereby empowering city administrations and stakeholders to craft policies and interventions that are not only data-driven but also deeply attuned to the voices of the urban populace.

Method

This study leveraged the rich potential of crowdsourced data to encapsulate urban opinion, drawing from diverse sources like social media and smart city platforms. The investigative focus of this study was on Indonesian urban centers: Jakarta. Employing an innovative web crawling methodology, the approach involved the strategic use of keywords that echo the social, economic, and environmental themes established (Ependi, Rochim, et al. 2023b), thereby ensuring a thorough and nuanced exploration. To guarantee the data's accuracy and contextual relevance, a geo-location filtering mechanism was employed, concentrating the data gathering efforts on these specific urban landscapes. This methodical approach culminated in the collection of a substantial dataset comprising 2,931 entries, each intricately aligned with the overarching theme of urban sustainability, spanning societal, economic, and environmental aspects. In the analytical phase, an annotation process was undertaken, with each piece of data being categorized as positive, negative, or neutral, through the InSet lexicon technique (Ependi, Aliya, and Wibowo 2023). This was done following well-defined weighting procedures, as shown in Figure 1.

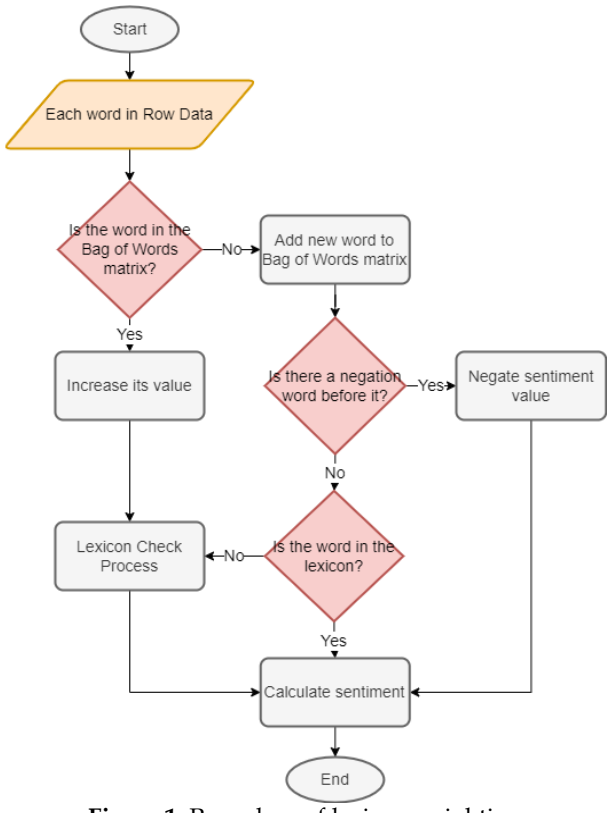

Figure 1. Procedure of lexicon weighting

Figure 1 delineates the procedure of sentiment weights: values falling below zero are deemed negative, while those exceeding zero are considered positive. A zero value is classified as neutral. Upon examining 2,931 rows of data, a conspicuous predominance of positive sentiments is observed, spanning across social, economic, and environmental dimensions. This trend is most evident, with over 600 rows affirming this pattern, as shown in Figure 2.

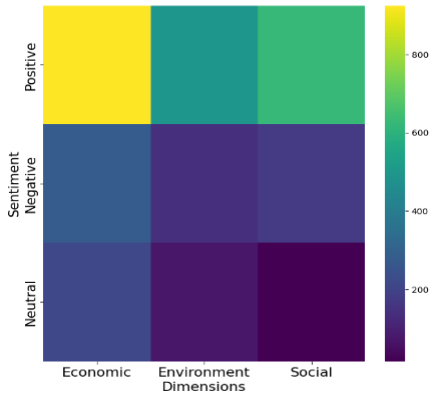

Figure 2. Data distribution for each dimension and sentiment

This investigation harnesses the power of machine learning to adeptly navigate the complexities of unbalanced data in the realm of urban opinion classification, leveraging insights from crowdsourced data. Figure 3 unveils the intricately designed machine learning architecture, where Random Oversampler (ROS) and Random Forest (RF) are employed as the cornerstone procedures in the intricate process of research resolution. Considering the intricate architecture exhibited in Figure 3, a comprehensive explanation unfolds as follows.

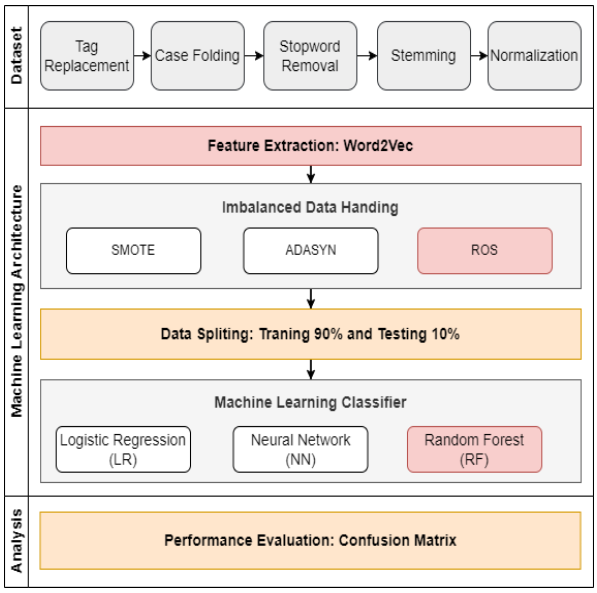

Figure 3. Proposed machine learning architecture ROS-RF

Dataset Handling

The handling of the dataset focuses on five key aspects to produce good quality data that impacts classification performance. These treatments include tag replacement, case folding, stopword removal, stemming, and normalization. (1) Tag replacement, a process of cleansing the row data from irrelevant and extraneous content such as entrance attachments, tabs, URLs, usernames, numbers, superfluous white spaces, and a variety of punctuation marks. (2) Case folding, a crucial step that ensures textual uniformity by converting all uppercase letters to their lowercase counterparts, thereby aligning words like 'Saya' and 'saya' under a single, standardized format. This is instrumental in mitigating variances in letter cases during the vectoring process (Macrohon et al., 2022). (3) The removal of stop-words, identified as words lacking substantial meaning, is conducted through the stopwords() function from NLTK, seamlessly integrated with the Sastrawi tool. (4) Stemming, a refined process of reducing words to their fundamental base form, is adeptly carried out using the Sastrawi library, effectively stripping words of their affixes such as suffixes and prefixes. (5) The process of normalization is particularly significant, tasked with the conversion of colloquial and non-standard language forms into a standardized version. This is indispensable considering the eclectic range of data sourced from Twitter, replete with slang expressions like 'bgt', 'dgn', 'slalu', 'gkmau', 'aq', 'yuuuk', 'sippp', etc. For this transformation, an extensive Indonesian dictionary comprising 17,321 texts is employed, ensuring comprehensive and accurate normalization of words (Ependi, Rochim, et al. 2023a).

Feature Extraction

This study adopts Word Embedding for the intricate process of feature extraction. Word Embedding, also referred to as word distribution representation, is an advanced technique employed for transforming textual data into a series of vector values. The design of these vector values is intricately aligned with the visual interpretation of the constituent words (Lopez et al., 2020). A significant aspect of this approach is the strategic arrangement of words within the vectors, which underscores the importance of semantic nuances and syntactical structure in the text. This method is prevalently utilized across various domains of text mining research, such as in conducting sentiment analysis and developing topic models (Abdi et al., 2019). Prominent word embedding formats include the vector arrangements exemplified by word2vec and fastText (Sastrawan et al., 2022). Despite the availability of these options, word2vec has been distinctively chosen as the embedding technique for this study, owing to its specific suitability and effectiveness.

Imbalanced Data Handling

The dataset, primed for the ensuing classification stage, confronted a formidable obstacle during the data collection process. The distribution of classes across multiple dimensions exhibited a marked imbalance, culminating in a significantly skewed and unequal distribution of samples. Such an imbalance had a pronounced impact on the classification's accuracy, frequently causing a skewness towards the more prevalent class.

The dilemma of an imbalanced data ratio is a well-known challenge, capable of substantially undermining the efficiency of classification algorithms if left unchecked (Tallo and Musdholifah 2018). To handle this obstacle, a variety of strategies were employed, prominently featuring data-level solutions like oversampling. For the purposes of this study, the random oversampler (ROS) was selected as the method for handling imbalanced data, with the procedures outlined in Figure 4.

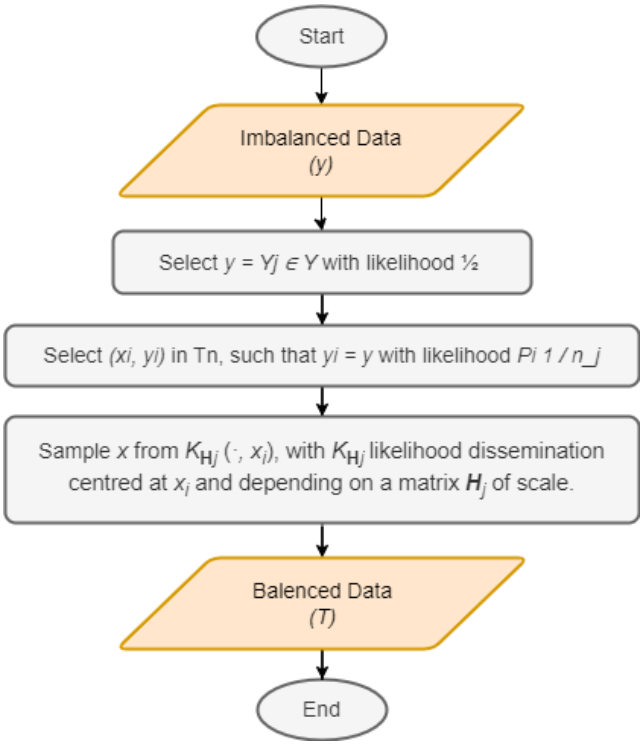

Figure 4. ROS procedure for handling imbalanced data

Data Spliting

In this study, the deliberate process of data splitting was employed, targeting enhanced accuracy in classification outcomes and effectively mitigating issues related to imbalanced datasets. A calculated ratio of 90% for training and 10% for testing was meticulously applied. The intricate process of data division was adeptly managed using the sklearn.model selection library. Furthermore, the practice of cross-validation was incorporated, designating a segment of the dataset for the development of the predictive model and another segment for a comprehensive evaluation of its performance (Diniz 2022).

Machine Learning Classifier

This study introduces a classification technique utilizing Random Forest combined with hybrid imbalanced data through Random Over Sampling (ROS) (Zhong & Wang, 2023). The performance of the Random Forest classifier is compared with that of Logistic Regression and Neural Network models. A Random Forest, which consists of multiple decision trees, classifies new objects based on their attributes. Each tree independently makes a classification, effectively "voting" for a particular class. The final classification is determined by the majority vote of all trees in the forest (Zhang & Wang, 2021), as illustrated in Figure 5.

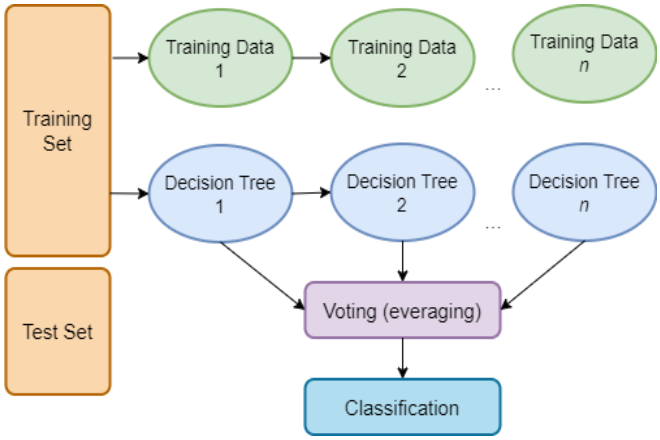

Figure 5. Random Forest Algoritm

The process of growing each tree in the forest involves the following steps: Sampling: From a training set with N cases, a random sample of N cases is taken. This sample serves as the training set for the tree. Variable Selection: Given M input variables, a number m (where m << M) is specified. At each node of the tree, m variables are randomly selected from the M variables, and the best split among these m variables is used to split the node. The value of m remains constant throughout the process. Tree Growth: Each tree is grown to its maximum possible extent without pruning.

Performance Evaluation

The training phase of the Accuracy Value (AV) model was marked by an exemplary demonstration of accuracy. Its performance was evaluated using a confusion matrix, which segmented sentiment into categories: positive, neutral, and negative. Integral metrics, including accuracy, precision, recall, and F-measure, were calculated based on this matrix (Meng et al., 2021). Accuracy represents the percentage of inputs that were correctly predicted by the model, typically

seen through a diminishing loss value. Precision is concerned with the model's capacity to precisely identify correct inputs, while recall quantifies the extent to which the model accurately identifies true positives. The f-measure serves as a harmonized average of both precision and recall. The methodologies for calculating these metrics - accuracy, precision, recall, and f-measure are systematically delineated in Equations (1), (2), (3), and (4).

$$\text{Accuracy} = \frac{(TP+TN)}{(TP + FP + FN + TN)} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F-Measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

Result and Discussion

Following the methodology outlined in Figure 3, this study's findings can be detailed as follows.

Experiment Results

The urban opinion classification process using the ROS-RF method begins with addressing imbalanced data, followed by the classification stage. Handling imbalanced data is crucial for ensuring accurate and reliable classification results. This study employed Random Over-Sampling (ROS) and compared its effectiveness against imbalanced data distribution, Adaptive Synthetic (ADASYN), and Synthetic Minority Over-sampling Technique (SMOTE).

In this study, the imbalanced data distribution consisted of 2042 rows for the positive class, 591 rows for the negative class, and 298 rows for the neutral class. When using the SMOTE approach, an equal distribution among positive, negative, and neutral classes was achieved, with each class containing 2042 rows. Similarly, ROS also provided an equal distribution, resulting in 2042 rows for each class. In contrast, the ADASYN method produced a varied distribution with 2080 rows for the positive class, 2042 rows for the negative class, and 1988 rows for the neutral class.

These findings underscore the differences in data handling approaches and their impact on class distribution, visually represented in Figure 6. By thoroughly analyzing these methods, this study demonstrates a comprehensive understanding of how each approach influences distribution and classification outcomes. The results suggest that while SMOTE and ROS achieve balanced distributions, ADASYN offers a more nuanced approach by adjusting the sample size based on the density of the minority class, potentially leading to more robust classification performance in certain scenarios.

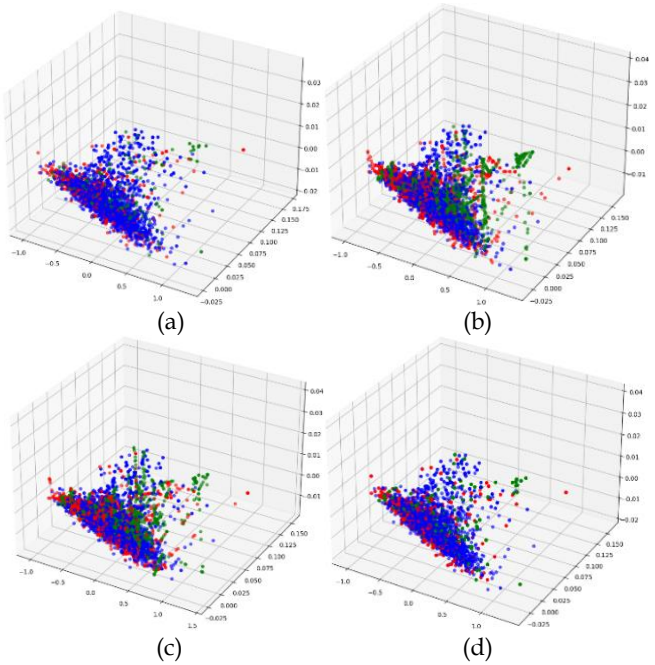

Figure 6. Feature distribusion (a) imbalanced, (b) SMOTE, (c) ADASYN, and (d) ROS

Following the methodology outlined in Figure 3, this study's findings can be detailed as follows. The performance evaluation of classification models using imbalanced data is presented in Table 1. Among the models, Random Forest exhibited the highest performance, achieving a precision of 0.84, recall of 0.81, and an F1-score of 0.78. In comparison, Logistic Regression had the lowest precision at 0.49 but a relatively high recall of 0.70, leading to an F1-score of 0.58. The Neural Network model displayed balanced precision and recall, both at 0.70, resulting in an F1-score of 0.60.

Table 1. Performa evaluation with imbalanced data

Model	precision	recall	f1-score
Logistic Regression	0.49	0.70	0.58
Neural Network	0.60	0.70	0.60
Random Forest	0.84	0.81	0.78

Table 2. Performa evaluation with SMOTE

Model	precision	recall	f1-score
Logistic Regression	0.47	0.45	0.45
Neural Network	0.57	0.56	0.56
Random Forest	0.96	0.96	0.96

Table 2 shows the performance evaluation after implementing the SMOTE technique. The Random Forest model's performance improved markedly, with precision, recall, and F1-score all reaching 0.96.

Conversely, both Logistic Regression and Neural Network models experienced a decline in performance relative to imbalanced data. Logistic Regression's F1-score dropped to 0.45, while the Neural Network's F1-score decreased to 0.56.

The results using the ADASYN technique are detailed in Table 3. The Random Forest model continued to surpass the other models, achieving precision, recall, and an F1-score of 0.95. The performance of Logistic Regression and Neural Network models remained relatively steady compared to SMOTE, with Logistic Regression showing an F1-score of 0.47 and the Neural Network model an F1-score of 0.54.

Table 3. Performa evaluation with ADASYN

Model	precision	recall	f1-score
Logistic Regression	0.48	0.47	0.47
Neural Network	0.54	0.55	0.54
Random Forest	0.95	0.95	0.95

Table 4. Performa evaluation with ROS

Model	precision	recall	f1-score
Logistic Regression	0.48	0.47	0.46
Neural Network	0.56	0.56	0.56
Random Forest	0.98	0.98	0.98

Table 4 outlines the performance metrics following the application of the ROS technique. The Random Forest model achieved the highest performance metrics across all evaluations, with precision, recall, and F1-score reaching 0.98. The Neural Network model showed a slight improvement with an F1-score of 0.56, whereas Logistic Regression maintained an F1-score of 0.46.

Across all resampling techniques, the Random Forest model consistently outperformed the Logistic Regression and Neural Network models. The performance of Random Forest was particularly noteworthy with ROS, achieving the highest scores across all metrics. This indicates that Random Forest is highly effective in handling imbalanced data, especially when augmented with ROS. In contrast, Logistic Regression and Neural Network models exhibited mixed results with different resampling techniques. Both models demonstrated decreased performance with SMOTE compared to imbalanced data, while their performance remained relatively consistent with ADASYN and ROS.

The result shows that resampling techniques can significantly impact the performance of classification models, with the combination of ROS and Random Forest providing the most substantial improvement in handling imbalanced datasets. These findings underscore the importance of selecting appropriate models and resampling methods to enhance classification performance in scenarios involving imbalanced data. The superior of ROS and Random Forest also show in testing data as shown in Figure 7.

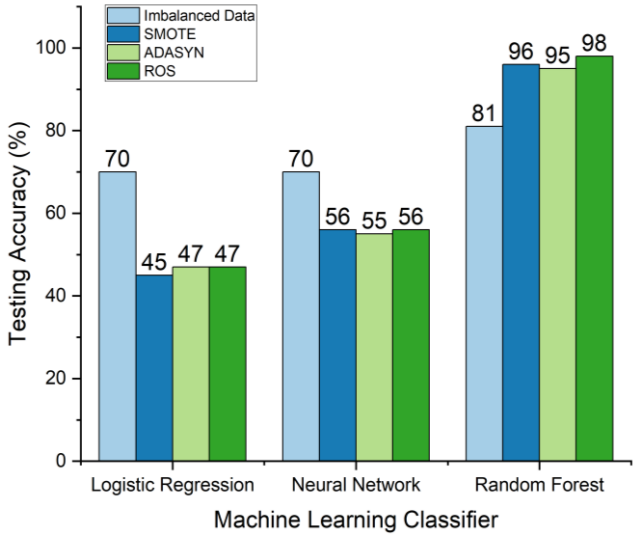

Figure 7. Testing Accuracy for each imbalanced data tecniques

Discussion

The urban opinion classification using the ROS-RF method underscores the critical importance of addressing imbalanced data to ensure accurate and reliable classification outcomes. This study meticulously compared the effectiveness of various resampling techniques, including Random Over-Sampling (ROS), Synthetic Minority Over-sampling Technique (SMOTE), and Adaptive Synthetic (ADASYN), in handling imbalanced datasets. The findings reveal significant insights into the impact of these techniques on class distribution and the subsequent performance of classification models.

The implementation of resampling techniques had a pronounced impact on model performance. When using the SMOTE technique, the Random Forest model's performance improved significantly, with all metrics (precision, recall, and F1-score) reaching 0.96. However, both Logistic Regression and Neural Network models experienced a decline in performance, with F1-scores dropping to 0.45 and 0.56, respectively. This suggests that while SMOTE effectively balances the data, it may not be universally beneficial for all classification models.

The ADASYN technique, detailed in Table 3, demonstrated a consistent performance for the Random Forest model, achieving precision, recall, and an F1-score of 0.95. Logistic Regression and Neural Network models showed steady performance compared to SMOTE, with F1-scores of 0.47 and 0.54, respectively. This indicates that ADASYN's nuanced sampling approach can maintain or slightly improve model

performance without the substantial fluctuations observed with SMOTE.

The ROS technique, resulted in the highest performance metrics for the Random Forest model, with precision, recall, and F1-score reaching 0.98. The Neural Network model showed a slight improvement with an F1-score of 0.56, while Logistic Regression maintained an F1-score of 0.46. These results underscore the effectiveness of ROS in conjunction with Random Forest, highlighting its superior ability to handle imbalanced data and enhance classification performance.

Across all resampling techniques, the Random Forest model consistently outperformed Logistic Regression and Neural Network models. The ROS-RF combination was particularly noteworthy, achieving the highest scores across all metrics and demonstrating its robustness in handling imbalanced datasets. In contrast, Logistic Regression and Neural Network models exhibited mixed results with different resampling techniques, with performance often declining or remaining steady compared to the imbalanced data baseline.

These findings emphasize the critical role of selecting appropriate resampling methods and classification models to optimize performance in the context of imbalanced data. The superior performance of the ROS-RF combination suggests that Random Forest, augmented with ROS, is a highly effective approach for urban opinion classification tasks involving imbalanced datasets. Future research could explore the application of these findings to other domains and investigate the potential for combining multiple resampling techniques to further enhance classification accuracy.

This study highlights the importance of addressing imbalanced data and the impact of different resampling techniques on classification performance. The ROS-RF method stands out as a particularly effective combination, offering substantial improvements in handling imbalanced datasets and achieving high classification accuracy. These insights contribute to a deeper understanding of resampling techniques and their role in enhancing model performance, providing a valuable foundation for future research and practical applications in the field of urban opinion classification.

Conclusion

This study successfully proposed a hybrid approach combining Random Over-Sampling (ROS) and Random Forest (RF) to handle imbalanced data in urban opinion classification. The ROS-RF method outperformed other techniques, such as Adaptive Synthetic (ADASYN) and Synthetic Minority Over-sampling Technique (SMOTE). In particular, the Random Forest model, when augmented with ROS, consistently displayed superior performance across various metrics, compared to Logistic Regression and Neural Network models. This study underscores the significant impact of resampling techniques on the performance of classification models, especially in imbalanced data scenarios. Furthermore, these findings highlight the need for careful selection of appropriate models and resampling methods to improve classification performance. Lastly, the robustness of the ROS and Random Forest combination was further confirmed through testing data, reinforcing its effectiveness.

Acknowledgments

We extend our heartfelt gratitude to Bina Darma University for their unwavering support, which has been instrumental in the successful completion of this research. We also express our sincere appreciation to the JPPIPA journal team for their willingness to review and publish this article.

Author Contributions

Conceptualization, U.E. and N.A.A.; methodology, U.E.; software, U.E.; validation, U.E. and N.A.A; formal analysis, U.E.; investigation, N.A.A.; resources, U.E.; data curation, N.A.A.; writing—original draft preparation, U.E.; writing—review and editing, U.E. and N.A.A.; visualization, U.E.; supervision, N.A.A.; project administration, U.E. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest

References

Abdi, A., Shamsuddin, S. M., Hasan, S., & Piran, J. (2019). Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion. *Information Processing & Management*, *56*(4), 1245-1259. https://doi.org/10.1016/j.ipm.2019.02.018.

Castelnovo, W., Misuraca, G., & Savoldelli, A. (2016). Smart cities governance: The need for a holistic approach to assessing urban participatory policy making. *Social Science Computer Review*, *34*(6), 724-739. https://doi.org/10.1177/0894439315611103.

Chen, H., Wu, L., Chen, J., Lu, W., & Ding, J. (2022). A comparative study of automated legal text classification using random forests and deep learning. *Information Processing & Management*, *59*(2), 102798. https://doi.org/10.1016/j.ipm.2021.102798.

Crooks, A., Pfoser, D., Jenkins, A., Croitoru, A.,

Stefanidis, A., Smith, D., ... & Lamprianidis, G. (2015). Crowdsourcing urban form and function. *International Journal of Geographical Information Science*, *29*(5), 720-741. https://doi.org/10.1080/13658816.2014.977905.

Diniz, M. A. (2022). Statistical methods for validation of predictive models. *Journal of Nuclear Cardiology*, *29*(6), 3248-3255. https://doi.org/10.1007/s12350-022-02994-7.

Ependi, U., Aliya, S., & Wibowo, A. (2023). Sentiment Analysis of Covid-19 Handling in Indonesia Based on Lexicon Weighting. *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, *12*(1), 76-82. https://doi.org/10.32736/sisfokom.v12i1.1615.

Ependi, U., Rochim, A. F., & Wibowo, A. (2023a). A Hybrid Sampling Approach for Improving the Classification of Imbalanced Data Using ROS and NCL Methods. *International Journal of Intelligent Engineering and Systems*, *16*(3), 345-361. https://doi.org/10.22266/ijies2023.0630.28.

Ependi, U., Rochim, A. F., & Wibowo, A. (2023b). An assessment model for sustainable cities using crowdsourced data based on general system theory: a design science methodology approach. *Smart Cities*, *6*(6), 3032-3059. https://doi.org/10.3390/smartcities6060136.

Ghahramani, M., Galle, N. J., Duarte, F., Ratti, C., & Pilla, F. (2021). Leveraging artificial intelligence to analyze citizens' opinions on urban green space. *City and Environment Interactions*, *10*, 100058. https://doi.org/10.1016/j.cacint.2021.100058.

Jalal, N., Mehmood, A., Choi, G. S., & Ashraf, I. (2022). A novel improved random forest for text classification using feature ranking and optimal number of trees. *Journal of King Saud University-Computer and Information Sciences*, *34*(6), 2733-2742. https://doi.org/10.1016/j.jksuci.2022.03.012

Joia, L. A., & Kuhl, A. (2019). Smart city for development: A conceptual model for developing countries. In *International conference on social implications of computers in developing countries*, 203-214. https://doi.org/10.1007/978-3-030-19115-3_17

Kourtzanidis, K., Angelakoglou, K., Apostolopoulos, V., Giourka, P., & Nikolopoulos, N. (2021). Assessing impact, performance and sustainability potential of smart city projects: Towards a case agnostic evaluation framework. *Sustainability*, *13*(13), 7395. https://doi.org/10.3390/su13137395.

Long, Y., & Liu, L. (2016). Transformations of urban studies and planning in the big/open data era: A review. *International Journal of Image and Data Fusion*, *7*(4), 295-308. https://doi.org/10.1080/19479832.2016.1215355.

Lopez, W., Merlino, J., & Rodriguez-Bocca, P. (2020). Learning semantic information from Internet Domain Names using word embeddings. *Engineering Applications of Artificial Intelligence*, *94*, 103823. https://doi.org/10.1016/j.engappai.2020.103823.

Macrohon, J. J. E., Villavicencio, C. N., Inbaraj, X. A., & Jeng, J. H. (2022). A semi-supervised approach to sentiment analysis of tweets during the 2022 Philippine presidential election. *Information*, *13*(10), 484. https://doi.org/10.3390/info13100484.

Meng, F., Cheng, W., & Wang, J. (2021). Semi-supervised software defect prediction model based on tri-training. *KSII Transactions on Internet & Information Systems*, *15*(11), 40–42. https://doi.org/10.3837/tiis.2021.11.009.

Salles, T., Gonçalves, M., Rodrigues, V., & Rocha, L. (2018). Improving random forests by neighborhood projection for effective text classification. *Information Systems*, *77*, 1-21. https://doi.org/10.1016/j.is.2018.05.006.

Sastrawan, I. K., Bayupati, I. P. A., & Arsa, D. M. S. (2022). Detection of fake news using deep learning CNN–RNN based methods. *ICT express*, *8*(3), 396-408. https://doi.org/10.1016/j.icte.2021.10.003.

Tallo, T. E., & Musdholifah, A. (2018). The implementation of genetic algorithm in smote (synthetic minority oversampling technique) for handling imbalanced dataset problem. In *2018 4th international conference on science and technology (ICST)*, 1-4. https://doi.org/10.1109/ICSTC.2018.8528591

Tan, S. Y., & Taeihagh, A. (2020). Smart city governance in developing countries: A systematic literature review. *sustainability*, *12*(3), 899. https://doi.org/10.3390/su12030899.

Tomor, Z., Meijer, A., Michels, A., & Geertman, S. (2019). Smart governance for sustainable cities: Findings from a systematic literature review. *Journal of urban technology*, *26*(4), 3-27. https://doi.org/10.1080/10630732.2019.1651178.

Viale Pereira, G., Cunha, M. A., Lampoltshammer, T. J., Parycek, P., & Testa, M. G. (2017). Increasing collaboration and participation in smart city governance: A cross-case analysis of smart city initiatives. *Information Technology for Development*, *23*(3), 526-553. https://doi.org/10.1080/02681102.2017.1353946.

Wang, W., Zhu, X., Lu, P., Zhao, Y., Chen, Y., & Zhang, S. (2024). Spatio-temporal evolution of public opinion on urban flooding: Case study of the 7.20 Henan extreme flood event. *International Journal of Disaster Risk Reduction*, *100*, 104175. https://doi.org/10.1016/j.ijdrr.2023.104175.

Webster, C. W. R., & Leleux, C. (2018). Smart governance: Opportunities for technologically-mediated citizen co-production. *Information*

Polity, *23*(1), 95-110. https://doi.org/10.3233/IP-170065.

Zhang, X., & Wang, M. (2021). Weighted random forest algorithm based on bayesian algorithm. In *Journal of Physics: Conference Series, 1924*(1). https://doi.org/10.1088/1742-6596/1924/1/012006.

Zhong, Yu, & Huiling Wang. (2023). Internet Financial Credit Scoring Models Based on Deep Forest and Resampling Methods. *IEEE Access, 11*, 8689–8700. https://doi.org/10.1109/ACCESS.2023.3239889.