Received: February 28, 2023. Revised: March 28, 2023.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A Hybrid Sampling Approach for Improving the Classification of Imbalanced Data Using ROS and NCL Methods

Usman Ependi[1,2]* Adian Fatchur Rochim[3] Adi Wibowo[4]

[1] *Doctoral Program of Information Systems, Universitas Diponegoro, Semarang, Indonesia*
[2] *Faculty of Science and Technology, Universitas Bina Darma, Palembang, Indonesia*
[3] *Faculty of Engineering, Universitas Diponegoro, Semarang, Indonesia*
[4]*Faculty of Science and Mathematics, Universitas Diponegoro, Semarang, Indonesia*
* Corresponding author's Email: u.ependi@binadarma.ac.id

Abstract: This research presents a novel hybrid sampling technique, implemented at the data level, to effectively address imbalanced and noisy data in classification processes. The proposed technique expertly combines two established methods, namely, the random over sampling (ROS) and neighbourhood cleaning rule (NCL) approaches, to tackle imbalance and noise issues, respectively. The study carried out an empirical evaluation of the proposed approach using crowdsourced text data that primarily emphasized the triple bottom line (TBL) dimension of a smart social, economic, and environmental city. The study used the long short-term memory (LSTM), convolutional neural networks (CNN), and CNN-LSTM classification models to validate the efficacy of the proposed hybrid sampling technique and compare its performance with other existing approaches, including ROS oversampling, NCL undersampling, synthetic minority over sampling & tomek links (SMOTE-Tomek), and synthetic minority over-sampling and edited nearest neighbours (SMOTE-ENN) hybrid sampling. The results are impressive, with the ROS-NCL hybrid sampling technique achieving high accuracy rates across all three classification models, at 97.71%, 98.01%, and 98.11%, respectively. This approach provides a robust and effective solution for handling impure data and holds great promise in identifying complex data patterns in real-world classification problems.

Keywords: Citizen opinion, Smart city dimension, Imbalanced data, Hybrid sampling, ROS-NCL.

1. Introduction

The classification process plays a crucial role in identifying various patterns in data. It is frequently utilized to address various challenges, such as identifying fraudulent transactions [1], diagnosing diseases [2], and detecting faults in air handling units [3]. A common issue faced during this process is imbalanced data, where one class has more significant features compared to other groups. This results in the majority-negative and minority-positive classes, where the former has a larger number of data points and the latter has fewer [4]. Given the importance of the minority class in classification, the handling of imbalanced data has become a challenging task in both machine and deep learning, as incorrect predictions may occur. Therefore, addressing imbalanced data becomes a crucial task in the field of classification.

These learning methods are widely employed to identify straightforward and suboptimal classification boundaries when organizing data with imbalanced class problems. The prevalence of this condition demonstrates that minority classes are often misclassified [5]. For instance, the classification of unbalanced data into the dimensions of smart cities aligns with the economical majority and social minority variables, with ratios of 80% and 20% respectively. When minority data is categorized into the negative class, high accuracy is observed, however, this condition becomes unreliable due to the random detection of the small data class [6]. As a result, machine or deep learning classification methods are limited in their ability to handle unbalanced data types.

To overcome this challenge, several approaches have been implemented, namely the data-level, algorithm-level, and hybrid-level approaches [7]. The data-level approach tackles the issue through the use of sampling methods, while the algorithm-level approach achieves similar results through one-class, cost-sensitive, or ensemble techniques [8]. The hybrid-level approach, on the other hand, integrates the data and algorithm levels through a mixed expert system. Among these approaches, the data-level approach is the most commonly used method to address imbalanced information, due to its advantage of increasing data validity and reducing training errors [9]. The data-level approach can also be advanced through the use of hybrid sampling, which balances the data and reduces noise by combining oversampling and undersampling [10].

The algorithmic-level approach aims to enhance classification results and reduce data imbalance through optimization techniques. Several studies have investigated algorithm optimization, with notable examples including the improved harris hawk optimization and opposition-based learning (IHHOOBL) algorithm, which is specifically designed to detect communities in social networks [11]; Several algorithms have been developed to address optimization problems, including the quantum-learning, gaussian, cauchy, and tunicate swarm (QLGCTSA) algorithm, which is a general-purpose algorithm [12]; the slime mould algorithm (SMA), designed to simulate biological wave optimization [13]; the sparrow search algorithm (SSA), developed specifically for optimization problems [14]; the tree seed algorithm (TSA), which identifies tree and seed relationships for optimization [15]; and QC-inspired metaheuristic algorithms, which aim to solve numerical optimization problems [16]. Although previous studies have shown that optimizing these algorithms, either through feature selection [17] or for improved accuracy, does not adequately address the challenge of imbalanced data arising from oversampling or undersampling.

On the contrary, balancing the distribution of data is a crucial aspect of classification since it has a significant impact on the overall performance of the classifier. Several previous studies have explored the challenge of imbalanced data by utilizing different hybrid sampling techniques, such as monte carlo mega-trend-diffusion (MCMTD), synthetic minority oversampling technique & reverse k-nearest neighbors (SMOTE-RkNN), hybrid of data-level & algorithmic-level (HybridDA), combined synthetic oversampling & undersampling technique (CSMOUTE), and multi class combined cleaning and resampling (MC-CCR).

The MCMTD approach employs gaussian fuzzy and mega-trend diffusion techniques to oversample the data and generate new instances of the minority class, respectively [18]. This method focuses on resolving imbalanced data problems in binary classification, generating virtual samples for the minority class, and is suitable for handling imbalanced data for numerical data types. The SMOTE-RkNN method integrates SMOTE and rough-set techniques to oversample and control new instances, respectively [19]. SMOTE-RkNN addresses imbalanced data by identifying noise based on probability density instead of noisy neighborhoods when creating new samples. However, this process is time-consuming, and SMOTE-RkNN is more focused on handling imbalanced data for numerical data types in the context of binary classification.

HybridDA amalgamates SMOTE oversampling, random undersampling (RUS), and SVM optimization utilizing grid search [20]. This method combines both data level and algorithm level approaches, using the data level for generating samples and the algorithm level for optimization. HybridDA primarily focuses on handling imbalanced data with binary classes. CSMOUTE employs synthetic generation and removes instances of both minority and majority classes [21]. Addressing imbalanced data by oversampling the minority class and undersampling the majority class, CSMOUTE is more focused on binary classification with numerical data types. MC-CCR is an approach that aims to resolve imbalanced data problems by first cleaning the data and then oversampling [22]. This method is appropriate for numerical data types and imbalanced multi-class classification. However, its lack of stability may render it unsuitable for real-world imbalanced data classification.

In light of the presented approaches, it can be concluded that MCMTD, SMOTE-RkNN, HybridDA, and CSMOUTE are particularly suitable for addressing imbalanced data classification problems in the context of binary classification with numerical data types [10]. However, the MC-CCR approach is designed for multi-class classification, although it has only been tested with numerical data types. Its effectiveness with other types of data, such as images and text, remains untested. Furthermore, the MC-CCR approach may lead to unstable performance and is not suitable for addressing real-world imbalanced data classification challenges.

Imbalanced data challenges become increasingly complex when dealing with multi-class data because the relationships between classes become more intricate. The multi-class classification problem

involves multi-minorities or multi-majorities. In practice, these relationships can be even more complex. The challenges associated with imbalanced data classification are amplified in multi-class settings, where every additional class increases the complexity of the classification problem. Binary approaches for imbalanced multi-class data have limitations due to the intra-class complexity. In addition to overcoming imbalanced data, different data types present significant challenges that also affect classification performance. Numerical, image, and text data require different treatments and face challenges that can affect classification performance [22, 23].

Given the aforementioned challenges, this study aims to propose a method for overcoming imbalanced data in text classification using hybrid sampling for muli-class classifiction. The proposed method combines random over sampling (ROS) and the neighborhood cleaning rule (NCL) to balance the distribution of classes. ROS randomly duplicates the minority class and discards instances from the majority class [24], while NCL removes noise from data distribution of each class [25]. ROS technique is a simplistic and straightforward oversampling method that can be efficiently executed with minimum computational complexity, while simultaneously ensuring that no data points are discarded. ROS has demonstrated superior performance in solving binary and multiclass classification problems and has been extensively studied and proven to achieve optimal performance in real-world classification scenarios [26]. On the other hand, the NCL method is designed to enhance the data cleanliness of the majority class in imbalanced datasets by considering the quality of the removed data. Unlike ROS, NCL primarily focuses on data cleansing instead of class balancing in the training set [27]. Therefore, the fusion of ROS and NCL in the form of Hybrid Sampling can complement each other, leading to enhanced classification performance.

To validate the proposed method, the opinions of citizens regarding the triple bottom line (TBL) of smart social, economic, and environmental city dimensions were collected from various social media platforms, such as Twitter. The opinions were then classified using a deep learning algorithm incorporating a long short-term memory (LSTM) and a convolutional neural network (CNN). A thorough review was conducted to ensure the validity of the results, and it was found that ROS-NCL was not previously applied to multi-class imbalanced classification.

This study presents several valuable contributions. Firstly, it develops a specialized text dataset that is tailored to Indonesia's smart city dimensions. Secondly, it utilizes improved methods to balance the classes in the dataset. Thirdly, it employs a hybrid sampling technique that combines ROS and NCL to increase accuracy in the context of text classification. Lastly, it compares and evaluates the proposed method with other hybrid data-level algorithms. The paper is organized into four main sections: (1) Introduction, which discusses the problem of imbalanced data and the strengths of the proposed approach, (2) Methods, which outlines the methodology used in the study, (3) Results and Discussion, which presents and discusses the experimental results, and (4) Conclusion and future Work, which concludes the paper and highlights avenues for future research, emphasizing the strengths of the proposed hybrid sampling technique, ROS-NCL.

2. Methods

In this study, eight distinct stages were employed, as depicted in Fig. 1. These stages included: (1) Data collection, (2) Annotation, (3) Preprocessing, (4) Word embedding, (5) Data splitting, (6) Data balancing, (7) Implementation of a deep learning classifier, and (8) Performance evaluation. A detailed explanation of each stage is provided below.

3.1 Data collection

Prior to data collection, the development of several keywords was crucial in identifying the indicators for each dimension of the triple bottom line (TBL). The identification process involved conducting extensive literature reviews on smart city assessment [28], which served as a source of inspiration in shaping the indicators and keywords for the social, economic, and environmental dimensions.

The social dimensions were derived from several sources such as the sustainable development indicators [29], Lisbon ranking for smart sustainable cities [30], IESE cities in motion index [31], ITU-T Y.4903/L.1603 indicators [32], and sustainability perspectives indicators [33]. The economic dimensions were compiled from sources such as the smart city index master indicators survey [34], dimensions of the smart city vienna UT [35], sustainability perspectives indicators [33], characteristics of smart city indicators [36], Criteria set for evaluating smart cities [37], Lisbon ranking for smart sustainable cities [30], IESE cities in motion index [31], China smart city performance [38],

Table 1. Keywords for crawling

Dimension	Indicators	Keywords
Social (So)	equity	housing, property
	health	health, hospital, nutrition, sanitation, drinking water
	education	education, literacy, schooling
	security	security, unemployment, slavery, crime, criminality, peace, violence
	culture and equality	culture, equality, population, female workers
Economy (En)	innovation	entrepreneur, company, innovation, technology, industry
	income	income, salary, employment, poverty rate, finances
	infrastructure	infrastructure, cooperation, connections
	business opportunity	economic performance, consumption, trade, competitiveness, productivity
Environment (Em)	air	air, pollution, emissions, defilement, waste
	energy	renewable energy, electricity, green industry, solar energy
	public facilities	green space, parks, city parks, vehicles, public transport

Juniper analysis of smart city frameworks [39], smart city dimension [40], and smart city performance index [41].

As for the environmental dimensions, they were sourced from dimensions of the smart city vienna UT [35], criteria set for evaluating smart cities [37], assessing the effectiveness of smart transport [42], China smart city performance [38], ITU-T Y.4902/L.1602 indicator [43], smart city dimension [40], and city sustainability assessment [44]. The search data for indicators and keywords is presented in Table 1.

Following the creation of keywords for each dimension, social media data were obtained through crawling Twitter utilizing the Rapidminer application tool with an academic account. The search filters employed the generated keywords and location, specifically focusing on the provincial capital cities on the Indonesian island of Java designated as smart cities, consisting of Jakarta, Bandung, Semarang, and Surabaya. The location filter was restricted to a 20 km radius from each city's longitude and latitude. The crawling process took place from August 25th to October 25th, 2022, yielding 12,185 items of raw data related to social, economic, and environmental factors. After undergoing filtering and selection, 5,981 relevant data pieces pertaining to the smart city and its dimensions were obtained.

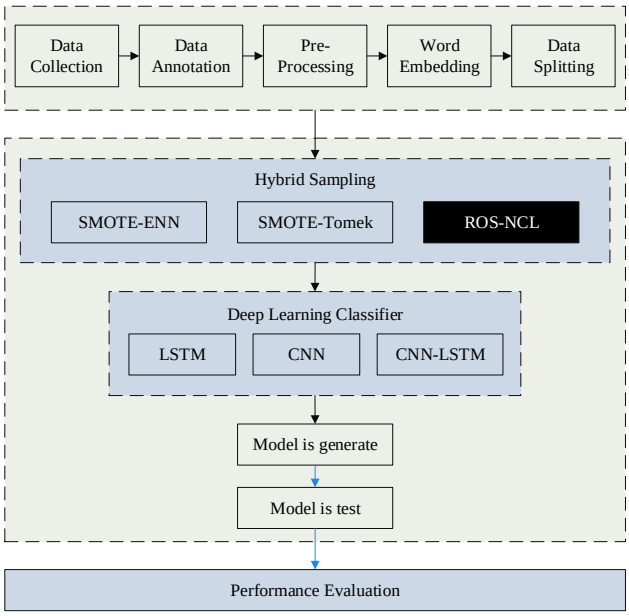

Figure. 1 Study framework

3.2 Data annotation

Following data collection and selection, all tweets were annotated with three labels: social, economic, and environmental. The social label focused on tweets related to equity, health, education, security, culture, and equality. The economic label prioritized tweets concerning innovation, income, infrastructure, and business opportunities. The environmental label represented tweets encompassing water, energy, and public facilities.

3.3 Preprocessing

The data cleaning and preprocessing technique employed natural language processing (NLP) to attain classification outcomes with high accuracy, as this processing method is crucial for the computer's understanding of data [45]. In this phase, various libraries, including Google Colab, nltk, pandas, spacy, and the Indonesian Sastrawi library were utilized for preprocessing. The Indonesian Sastrawi library particularly aided in converting the Indonesian language's word affix into its basic form [46].

The preprocessing process consisted of tag replacement, case folding, stopword removal, stemming, normalization, and tokenization. These components are further explained as follows: (1) Tag replacement involved removing unnecessary attachments on raw data, including entrance attachments, tabs, URLs, usernames, numbers, additional white spaces, exclamation points, question marks, special characters, and punctuations. (2) Case folding transformed all uppercase letters into lowercase forms to achieve general similarities among all the dataset characters. For instance, "Saya"

and "saya" were considered the same. This stage aimed to reduce the differences between lowercase, uppercase, and capital letters when vectoring [45]. (3) Stop-word removal eliminated meaningless words and was carried out using the stopwords() library provided by NLTK through the Sastrawi tool. (4) Stemming was performed to convert a word with an affix into its root form (base word) by removing affixes such as suffixes and prefixes. The stemming process was conducted using the Sastrawi library. (5) Normalization emphasized the conversion of a non-standard word into a common type or according to its spelling. This step was necessary due to the various data obtained from Twitter, which contained multiple slang such as "bgt," "dgn," "slalu," "gkmau," "aq," "yuuuk," "sippp," etc. To convert the normalization of words, a dictionary of Indonesian containing 17321 texts was used. (6) Tokenization involved breaking down sentences into pieces of words, punctuation marks, and other meaningful expressions. The word_tokenize() function provided by the nltk library was utilized in this process.

3.4 Word Embedding

Word Embedding (word distribution representation) is a technique for mapping texts into vector values. The form of these values is often arranged based on the visual interpretation of the words [47]. Additionally, word arrangement in vectors places a significant emphasis on semantic information and text syntax. This technique is widely used in various text mining studies, such as sentiment analysis and topic modeling [48]. Word embedding also commonly uses the arrangement of vector number shapes, such as word2vec and fastText [49]. Despite this, the word2vec or vectorizer were still selected as the preferred method of embedment for this study.

3.5 Data Splitting

The process of data splitting was utilized to achieve high accuracy in classification performance, as well as to mitigate the issue of imbalanced datasets. For this analysis, a splitting ratio of 90% for training and 10% for validation and testing was employed. The sklearn.model_selection library was utilized to facilitate the splitting process. Cross-validation was also applied, where one part of the data was used to develop the predictive model while the other was employed to evaluate its performance [50].

3.6 Imbalanced Approaches

The cleaned data was then referred to as the dataset and was ready for classification. However, during the data collection phase, the class distribution in each dimension was imbalanced, resulting in an uneven distribution of samples. This also impacted the classification performance, often leading to a bias towards the majority class. The issue of an imbalanced proportion of data is a common one, and if left unaddressed, can significantly reduce the performance of the classification algorithm [51]. To overcome this challenge, various approaches were adopted, including data-level techniques such as oversampling, undersampling, and feature selection, and algorithm-level techniques such as one-class, cost-sensitive, and ensemble methods [7]. The understanding of the imbalanced data ratio was achieved by calculating the majority and minority classes, as represented in Eq. (1), where $\sum Class_{mayority}$ and $\sum Class_{minority}$ refer to the respective classes and $Ratio\ (\rho)$ represents the imbalanced ratio between both classes.

$$Ratio\ (\rho) = \frac{\sum Class_{mayority}}{\sum Class_{minority}} \quad (1)$$

The ROS and NCL method were proposed as a solution to address imbalanced data in the classification of citizens' opinions on the smart city dimension. This approach focuses on the data level, utilizing a combination of oversampling and undersampling techniques. The ROS method effectively generates additional data to improve distribution and information during the training process, while the NCL method helps to eliminate data overlapping and noise. This led to the hypothesis that the ROS-NCL approach would result in optimal performance in classifying citizens' opinions on the smart city dimension. This approach was compared to other pre-existing techniques such as synthetic minority over-sampling and edited nearest neighbours (SMOTE-ENN) [51] and synthetic minority over sampling & tomek links (SMOTE-Tomek) [52]. This section outlines the implementation procedures of the ROS-NCL method in handling imbalanced data.

2.6.1. Random over sampling (ROS)

The ROS technique is a data-level approach aimed at addressing the issue of imbalanced data by increasing the number of minority classes. This is accomplished by randomly replicating instances to balance the majority classes [10]. The oversampling

technique is also implemented by examining the training data for one class, with a similar probability assigned to both Y_0 and Y_1. The ROS algorithm then rates new samples based on their neighbors and determines the sample width for Hj. Additionally, the selection of KH_j is based on a unimodal symmetric distribution. The following outlines the steps of the ROS algorithm [53]:

- Select $y = Y_j \in Y$ with likelihood $\frac{1}{2}$
- Select (x_i, y_i) in $\mathbf{T}_n$, such that $y_i = y$ with likelihood $p_i = \frac{1}{n_j}$
- Sample x from $K_{\mathbf{H}j}(\cdot, x_i)$, with $K_{\mathbf{H}j}$ likelihood dissemination centred at x_i and depending on a matrix $\boldsymbol{H}_j$ of scale.

2.6.2. Neighborhood cleaning rule (NCL)

The NCL algorithm, derived from the edited nearest neighbor rule (ENN), aims to improve imbalanced data by cleaning the majority classes. It is widely recognized as an effective undersampling technique, known for its ability to remove data while preserving high quality. NCL prioritizes information cleaning and removing noise from the training data over balancing class proportions [54]. The undersampling process starts by identifying the N1 sample and its three nearest neighbors in the training data. If N_1 belongs to the majority class and its classification result is inconsistent with the original group, NCL removes N_1. Conversely, if N_1 is a part of the minority class, the majority group will be removed as its neighbor [55]. The steps of the NCL algorithm are outlined as follows [54].

- Split data T into the class of interest C and the rest of data O
- Identify noise of data A_1 in O based on the ENN rule
- For each class, C_i in O is observed,
  - When (x Ci in 3-nearest neighbours of misclassified y C)
  - and ($| C_i | 0.5 \cdot | C |$) then $A_2 = \{x\}$ A_2
- Reduce data $S = T - (A_1 A_2)$

After explaining the working methods of ROS and NCL, we have combined them to create a hybrid sampling method called ROS-NCL. To begin, we initialize the imbalanced dataset as "y" and use the ROS method to oversample the minority class, which produces a balanced dataset. Then, we initialize the data as "T" and utilize the NCL method to eliminate noise from the dataset. The outcome of this two-step process is a clean and balanced dataset. A flowchart demonstrating the ROS-NCL hybrid sampling procedure is presented in Fig. 2.

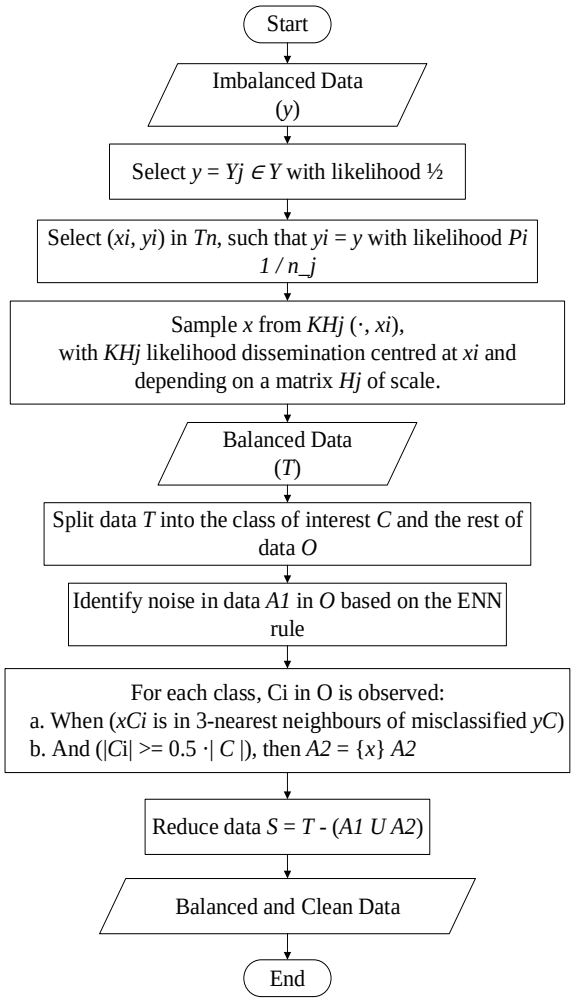

Figure. 2 The Proposed hybrid sampling ROS-NCL method

3.7 Deep learning classifier

The classification was performed using the LSTM and CNN algorithms. LSTM, which is often incorporated as part of RNN, has gained popularity for its use in text classification [56]. It overcomes the problem of gradient disappearance by incorporating a memory block in place of a self-connected hidden unit and consists of four components, including an input gate (i), which controls the size of new memory added, a forget gate (f), which determines the amount of memory forgotten, an output gate (o), which modulates the amount of resultant memory, and a cell activation vector (C), which consists of two components, including a portion of previous memory ($CT-1$) and a newly modulated type (CT) [57].

LSTM has been shown to perform better than RNN in text classification tasks [56]. According to [58], the algorithm performs well in various

classification scenarios, such as short, English-language, and news texts. Several studies have demonstrated its effectiveness, with [57] reporting an accuracy rate of 97% compared to other algorithms like SVM, Paragraph-Vec, and CNN-multi channels. These results were obtained from various public datasets, including the movie review sentence polarity dataset v1.0, IMDB, RT-2k, SST-1, SST-2, and TREC datasets. In addition, [59] found that LSTM outperformed RNN-Vanilla and GRU with an accuracy rate of 84% in classifying customer service texts in Indonesian. The algorithm has also shown good performance in classifying Indonesian hate speech and news texts [60], [61], as well as adult content on social media [62].

The convolutional neural network (CNN) is a type of neural network that consists of several layers, including the input, convolution, pooling, fully connected, and output layers. This network selects features by utilizing the convolution layer (CL) through a convolution kernel [63]. Its application to text classification was first introduced by [64] and produced impressive results. Since then, the algorithm has proven to be highly effective in various classification scenarios, including student learning needs [65], news story categorization [66], and Arabic text classification [67]. It has also shown its effectiveness in Indonesian text analysis, providing excellent results in classifying public opinions on the Covid-19 vaccine [68]. These findings demonstrate that the LSTM and CNN algorithms are well-suited for classifying citizens' opinions on the smart city dimension, utilizing crowdsourced data.

3.8 Performance evaluation

The accuracy value (AV) model demonstrated the highest level of accuracy during the training process. The predictions were obtained through the utilization of a confusion matrix, which categorized the dimensions of social, economic, and environmental TBL. The accuracy, precision, recall, and F-measure values were then derived from the same confusion matrix [69]. Accuracy reflects the proportion of inputs accurately predicted by the LSTM or CNN model and is indicated by a decrease in loss value. Precision focuses on the ratio of inputs accurately identified by the system, while recall calculates the proportion of inputs that were accurately recognized as true. The F-measure is an average of precision and recall. The formulas for calculating accuracy, precision, recall, and F-measure are presented in Eqs. (2), (3), (4), and (5), respectively.

$$\text{Accuracy} = \frac{(TP+TN)}{(TP + FP + FN + TN)} \quad (2)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

$$\text{F-Measure} = 2 \text{ x} \frac{\text{Precision x Recall}}{\text{Precision + Recall}} \quad (5)$$

TP (true positive) signifies the accurate prediction of real positive data, TN (true negative) represents the accurate prediction of real negative data, FP (false positive) denotes an erroneous prediction of positive data as positive, and FN (false negative) represents an erroneous prediction of positive data as negative.

3. Results and discussion

3.1 Experiment setup imbalanced data handling

In this study, datasets featuring social, economic, and environmental dimensions that had been preprocessed were trained using LSTM, CNN, and CNN-LSTM models. The training and validation & testing data were split into 90 and 10% respectively. The training process was carried out using the Keras library, with optimization algorithms such as Adam, Softmax, ReLU, and Sigmoid activation. The LSTM model was trained over 10 epochs, with a dropout of 0.2 and batch size of 64. The LSTM architecture featured 250 inputs and 100 outputs for word embedding, spatial dropout, and the LSTM model. The dense layer used 100 and 3 for input and output respectively. The CNN model was trained with a batch size of 64 and 10 epochs, utilizing 250 inputs and 100 outputs for word embedding, max pooling, and the CNN model. Additionally, the flattened data used 100 and 12500 inputs and outputs. The dense layer still utilized 100 and 3 as inputs and outputs. The CNN-LSTM model effectively combined the CNN and LSTM models as depicted in Fig. 3.

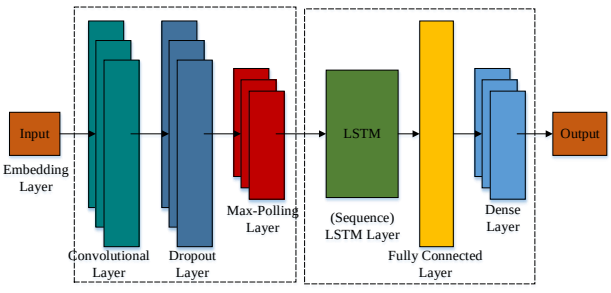

Figure. 3 CNN-LSTM architecture for training data

Table 2. Imbalanced ration

Minority	Majority	Imbalanced Ratio
So (1363)	En (3580)	2,63
Em (1038)	En (3580)	3,45

Table 3. Generated data from proposed hybrid sampling

Hybrid Sampling	Training	Validation & Testing
ROS	9666	1074
NCL	3731	415
ROS-NCL	9046	1006
SMOTE-Tomek	9630	1070
SMOTE-ENN	7334	815

Imbalanced data was handled while training the LSTM, CNN, and CNN-LSTM models using techniques such as ROS, NCL, SMOTE-Tomek, SMOTE-ENN, and ROS-NCL. The majority and minority classes were taken into account, with the economic dimension having 3580 rows, and the social and environmental dimensions having 1363 and 1038 rows respectively. The ratio of the majority to minority class was calculated using Eq. (1). The results showed that the social class required approximately 62% of the majority class data with a ratio of 2.63, while the environmental class required a higher percentage of 71% with a ratio of 3.45. These ratios are displayed in Table 2.

The data in Table 2 reveals the outcome of balancing imbalanced data through a combination of oversampling, undersampling, and hybrid sampling methods. As observed from the results, the ROS oversampling approach significantly boosted the data by 79%, yielding 10740 rows with a ratio of 1.8. In contrast, the NCL undersampling approach reduced the data by 30.7%, resulting in 4146 rows with a ratio of 0.7. On the other hand, the hybrid sampling approach elevated the data for all algorithms, with SMOTE-Tomek and SMOTE-ENN having the highest increase at 78.9% and 36.2% for 10700 and 8149 rows, respectively, resulting in ratios of 1.8 and 1.4. The ROS-NCL hybrid sampling approach also increased the data by 68.1% to 10052 rows, with a ratio of 1.7. Among all the imbalanced data techniques, ROS and SMOTE-Tomek achieved the highest sampling values, both with a ratio of 1.8. The results of oversampling, undersampling, and hybrid sampling methods are presented in Table 3.

The relationship between the impact of imbalanced techniques on sample generation and the number of features produced is direct. Each technique produces a distinct set of features, with SMOTE-Tomek generating the largest number at 102,230 features. NCL, on the other hand, produces the smallest number of features at 40,3559. ROS and ROS-NCL produce 99,886 and 94,364 features, respectively. SMOTE-ENN generates a larger number of features than NCL, with a total of 79,925.

The generated features showcase the total count of features produced by each imbalanced data technique used. Notably, the distribution of features varies based on the classification label, namely economic, social, or environmental, and the technique employed for balancing the data, as shown in Fig. 4. For instance, ROS allocates 35% of its features to the economic class (35,310 features), 33% to the social class (32,825 features), and 32% to the environmental class (31,751 features). On the other hand, NCL distributes 65% of its features to the economic class (26,617 features), 12% to the social class (4,999 features), and 23% to the environmental class (9,343 features). ROS-NCL, in contrast, allocates 37% of its features to the economic class (35,242 features), 31% to the social class (29,438 features), and 32% to the environmental class (29,684 features). Additionally, SMOTE-Tomek assigns 34% of its features to the economic class (35,310 features), 33% to the social class (33,502 features), and 33% to the environmental class (33,417 features). Lastly, SMOTE-ENN distributes 32% of its features to the economic class (25,516 features), 33% to the social class (26,736 features), and 35% to the environmental class (27,673 features). Based on these results, it can be concluded that all techniques result in a balanced distribution of features across classes. However, NCL still demonstrates an imbalance of 50% in the quantity of features due to its focus on reducing noisy rows. Similarly, the difference between the majority and minority class features in ROS-NCL reaches 6%. Fig. 4 illustrates the distribution of features for each imbalanced approach.

3.2 Performance exploration

The results of the data training approach, aimed at identifying the best classification model through the utilization of LSTM, CNN, and CNN-LSTM, are presented in Table 3. The training scenarios were designed to consider various sampling methods, including ROS, NCL, ROS-NCL, SMOTE-Tomek, and SMOTE-ENN. The use of LSTM showed a significant difference in the results between the sampling approaches during the first epoch, with SMOTE-Tomek and NCL achieving the lowest and highest accuracy levels of 45.6% and 60.9%, respectively. The second epoch saw NCL and ROS-NCL recording the lowest and highest increases in accuracy, with levels of 12.7% and 29%, respectively. NCL reached an accuracy level of 73.6%, while ROS-NCL achieved 88.2% accuracy. SMOTE-Tomek remained at the lowest accuracy level. During

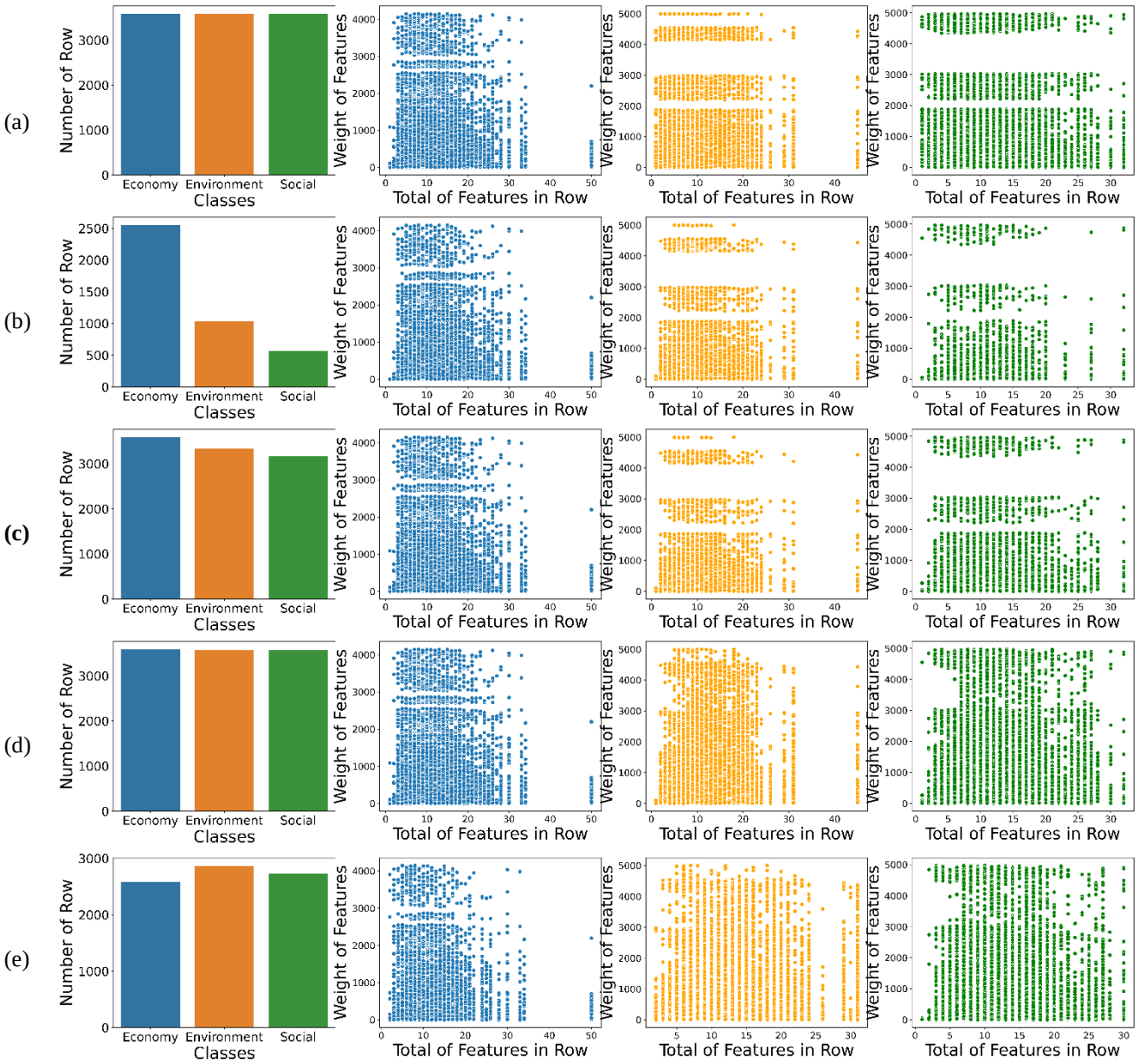

Figure. 4 Class distribution and Features Weight Based on Proposed Hybrid Sampling: (a) ROS, (b) NCL, (c) ROS-NCL, (d) SMOTE-Tomek, and (e) SMOTE-ENN

the third epoch, all sampling methods (oversampling, undersampling, and hybrid sampling) saw an increase in accuracy, ranging from 5% to 12%. The highest increase in accuracy was observed in SMOTE-ENN, while the lowest was seen in ROS. The best accuracy was recorded by ROS-NCL at 94.5%, while the worst was achieved by SMOTE-Tomek at 79%. ROS-NCL recorded a significant increase of 6.3%, while SMOTE-Tomek recorded a rise of 11%. NCL achieved an accuracy of 82.8%, with an increase of 9.2%, while SMOTE-ENN reached the same accuracy level with a similar increase. In the fourth epoch, the training process saw only a minor increase in accuracy, ranging from 1% to 7%. ROS recorded the lowest increase, while NCL recorded the highest. ROS-NCL achieved the highest accuracy at 96.7%, with a 2.2% increase. From the fifth to the tenth epochs, the training process saw a stagnation in accuracy increase, which was only observed in NCL, SMOTE-Tomek, and SMOTE-ENN, with increases ranging from 2.5% to 3.5%. The best accuracy levels exchanged between the first and tenth epochs in NCL and ROS-NCL, respectively. The lowest accuracy levels were consistently recorded in SMOTE-Tomek, both in the first and tenth epochs. However, the gap in accuracy between the beginning and end of the epochs decreased from 15.3% to 5.6%. A detailed representation of the LSTM model training process can be found in Fig. 5 (a).

The implementation of CNN in data training showed similarities with LSTM, however, to attain a 90% accuracy in handling imbalanced data, the best results were achieved at the fourth epoch, as depicted in Fig. 5 (b). The first epoch showed a range of accuracy levels, with SMOTE-Tomek recording the lowest at 43.4%, while NCL achieved the highest at 59.4%. The second epoch displayed a significant improvement, with increases up to 33.8% for ROS-NCL and a decrease of 15.5% for NCL. The trend continued in the third epoch, with rises of 5.2% for ROS-NCL and 15.6% for SMOTE-Tomek, and the highest and lowest accuracy levels recorded at 93% and 68% for ROS-NCL and SMOTE-Tomek, respectively.

At the fourth epoch, ROS-NCL recorded the highest accuracy of 98.8%, with NCL having the lowest. During the final epochs (5th to 10th), a decline in accuracy was observed for all approaches, with the highest increase of 4% for NCL, SMOTE-Tomek, and SMOTE-ENN. The best results were achieved at the end of the training, with ROS-NCL attaining the highest accuracy of 99.5% and ROS the lowest at 98%. These results showcase the dynamic nature of accuracy levels throughout the training process, with NCL starting with the highest accuracy and SMOTE-Tomek the lowest, but with ROS-NCL and ROS ending with the highest and lowest accuracy levels, respectively.

Based on the results presented in Fig. 5 (c), CNN-LSTM was determined that an accuracy rate above 90% was only achieved during the fifth epoch. During the early stages of the training process, the accuracy rates for SMOTE-Tomek and SMOTE-ENN were in the 40s, while ROS, NCL, and ROS-NCL recorded accuracy rates in the 60s. This highlights the significant variability in accuracy levels when compared to training LSTM and CNN models. The second epoch saw the largest increase in accuracy with ROS-NCL recording a 32.2% improvement. In the third epoch, SMOTE-ENN recorded a substantial increase of 86.9%, while ROS-NCL achieved the highest accuracy rate at 96.9%. During the fifth and sixth epochs, there was a minimal increase in accuracy for all imbalanced approaches, with increases ranging from 1% to 7%. The seventh to tenth epochs also saw an average increase in accuracy of less than 1%. In contrast, the training process using the CNN-LSTM model maintained consistent accuracy levels from the second to the tenth epoch, with ROS-NCL recording the highest accuracy and SMOTE-Tomek the lowest. Fig. 5 were used to train a model that would identify citizens' opinions on various dimensions of a smart city. However, the validity of the model must be confirmed to ensure accurate identification of social, economic, and environmental opinions. A significant difference of up to 20% was observed between the training and validation results in Figs. 6 and 7 for all classifiers (LSTM, CNN, and CNN-LSTM). This indicates that the model was only able to recognize data based on its training and was unable to accurately identify the validation information.

ROS was found to outperform SMOTE-Tomek and SMOTE-ENN, demonstrating that the difference between training and validation varied for each classifier. The F1 scores showed differences of 4%, 10%, and 11% for LSTM, CNN, and CNN-LSTM, respectively, as shown in Fig. 6. Meanwhile, the accuracy difference between training and validation was 2% for CNN and CNN-LSTM and 4% for LSTM. This result suggests that the LSTM, CNN, and CNN-LSTM models with ROS have not been optimized for accurately identifying citizens' opinions on smart city dimensions. The NCL approach was found to have a better f1 score gap compared to SMOTE-Tomek, SMOTE-ENN, and ROS, with differences of 4% and 15% observed for CNN/CNN-LSTM and LSTM, respectively. The accuracy gap between training and validation was also relatively high at 9% for all classifiers. This result indicates that the training data was not optimal for accurately identifying citizens' opinions using the LSTM, CNN, and CNN-LSTM models with the NCL approach. The model was unable to fully recognize the validation data.

In contrast, the ROS-NCL approach showed stability between training and validation for all classifiers, with a difference of less than 1% for CNN and CNN-LSTM and no difference for LSTM. This suggests that the resulting model was capable of accurately recognizing validation data based on its training information, resulting in a well-optimized model. As a result, the model is capable of accurately identifying citizens' opinions on smart city dimensions.

After the development of the model, rigorous validation and testing were performed to evaluate the accuracy of each approach for addressing imbalanced data, including the methods of ROS, NCL, ROS-NCL, SMOTE-Tomek, and SMOTE-ENN. The tests were conducted using 10% of the data, as depicted in Table 3, and the results were analyzed through the use of a confusion matrix. Figs. 7 and 8 depicted a correlation between the validation and testing accuracy, suggesting that there were no significant differences in the accuracy of each of the imbalanced data approaches for the LSTM, CNN, and CNN-LSTM models. However, variations were noticed when comparing the results from the training, validation, and testing phases. The methods of ROS,

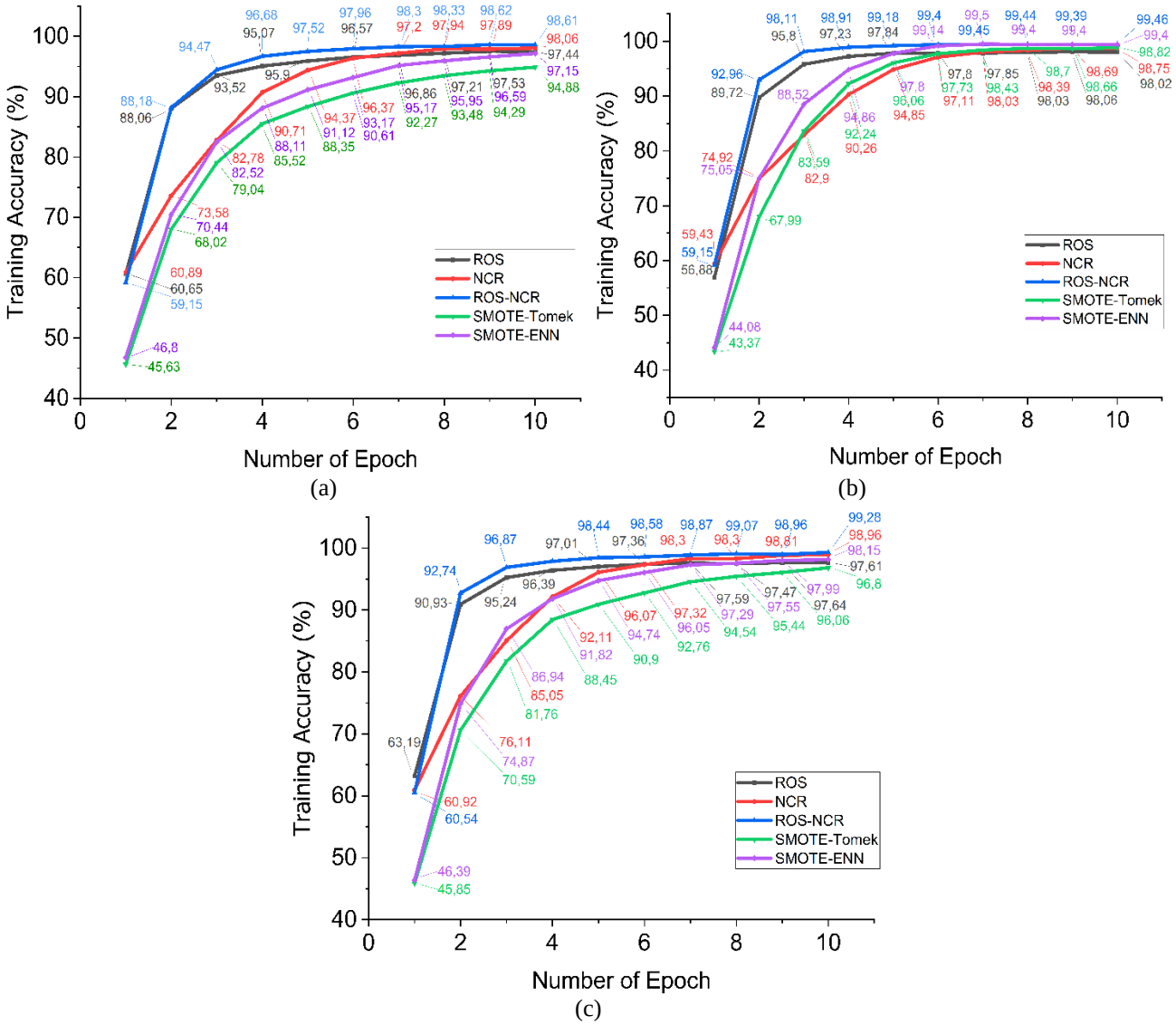

Figure. 5 Training accuracy of proposed hybrid sampling of ROS-NCL: (a) LSTM, (b) CNN, and (c) CNN-LSTM

NCL, SMOTE-Tomek, and SMOTE-ENN showed inconsistencies when comparing the training, validation, and testing results for all classifier models, but ROS-NCL demonstrated a critical alignment between these analytical components. The accuracy testing for ROS-NCL resulted in 97.71%, 98.01%, and 98.11% for the LSTM, CNN, and CNN-LSTM models, respectively. These findings demonstrate that ROS-NCL was the most effective approach in handling imbalanced data and classifying citizens' opinions into the dimensions of a smart city with precision.

3.3 Discussion

The findings demonstrate that each approach significantly impacted the classification performance, whether through oversampling, undersampling, or hybrid sampling. Among these methods, the CNN-LSTM model achieved the highest accuracy of 95.7% when ROS was used for oversampling. However, it was observed that ROS exhibited instability between the training and testing phases. On the other hand, NCL was implemented for undersampling, and the CNN model achieved the best performance with an accuracy of 90.36%. Nevertheless, like ROS, this classifier was also unstable. The best overall performance was obtained by using hybrid sampling, particularly ROS-NCL, indicating that both ROS and NCL play crucial roles in hybrid conditions. ROS increased the sample size, while NCL eliminated noise from the generated classes.

The performance of SMOTE-Tomek and SMOTE-ENN did not reach their maximum potential for all classification models when compared to hybrid sampling. The results showed that SMOTE-Tomek achieved the highest performance in the CNN-LSTM model at 74.65%, whereas SMOTE-ENN

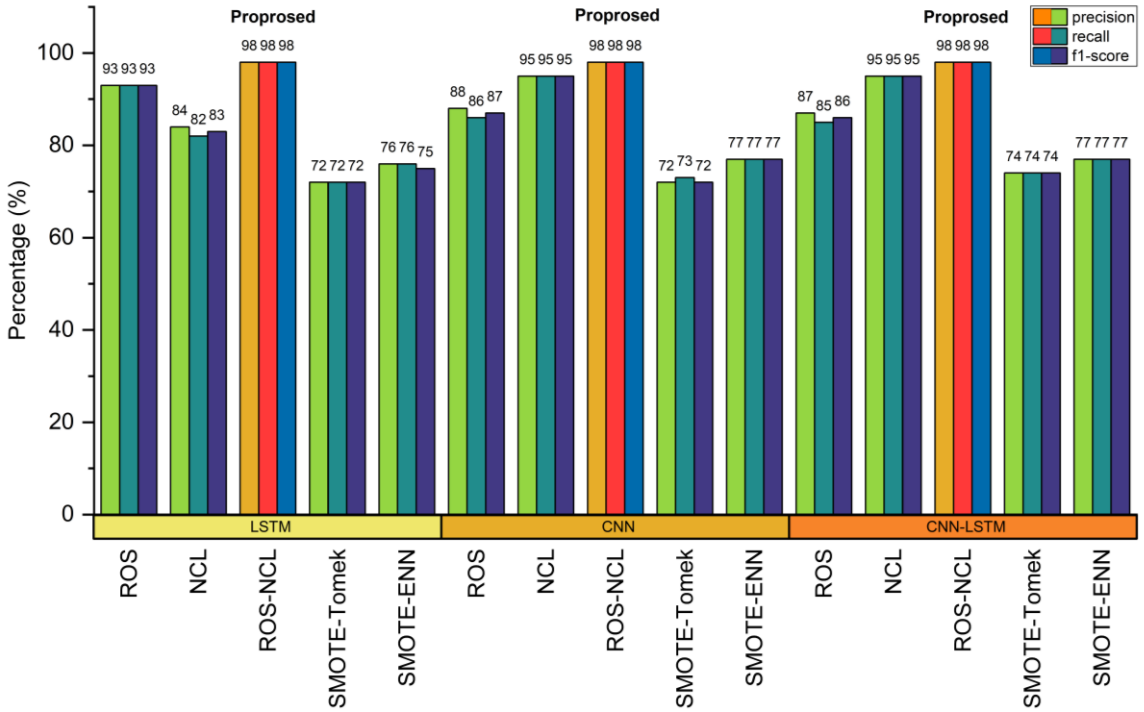

Figure. 6 Average precision, recall, and F1 score

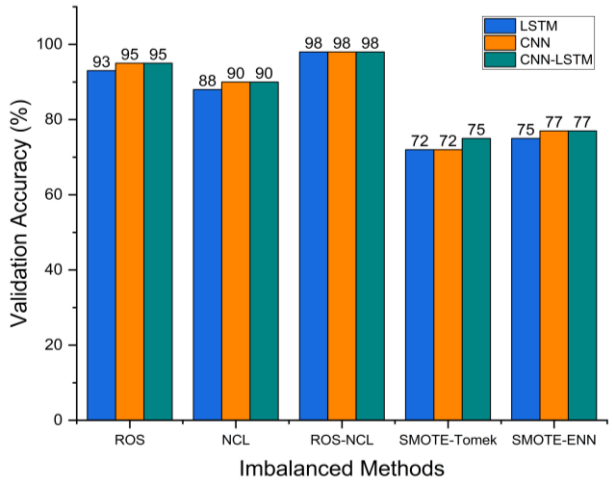

Figure. 7 Validation of the accuracy

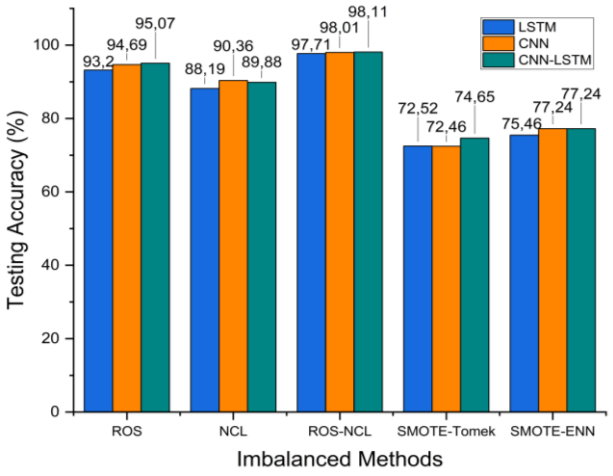

Figure. 8 Testing accuracy

demonstrated the best performance in the CNN and CNN-LSTM models at 77.24%. However, when compared to hybrid sampling, all classifiers, including ROS-NCL, SMOTE-Tomek, and SMOTE-ENN, exhibited significantly different performances. The lower performance of SMOTE-Tomek and SMOTE-ENN could be due to their suboptimal noise removal process, which was similar to ROS-NCL in terms of the percentage of training data. As such, based on the findings, ROS-NCL appears to be a superior option for improving imbalanced data and enhancing the performance of classifying citizen opinions regarding smart city dimensions. Additionally, the results emphasize the importance of addressing issues related to crowdsourced data [70].

An investigation was conducted to examine the effectiveness of the proposed hybrid sampling approach. This involved a comparative analysis of existing techniques, and Table 4 showed variations in class characteristics and data attributes. However, all of the hybrid sampling techniques, including MCMTD, SMOTE-RkNN, HybridDA, CSMOUTE, and MC-CCR, produced strong classification performances when applied at the data level. This included ROS-NCL, which demonstrated the efficacy of the proposed approach for addressing imbalanced data.

MCMTD hybrid sampling involves several stages to address imbalanced data, including establishing a forecasting model, determining attribute ranges, and generating virtual samples. During the process of

generating virtual samples, MCMTD requires the declaration of the number of samples to be produced. Experimental results using two datasets, namely multi-layer ceramic capacitors (MLCC) and purified terephthalic acid (PTA), showed a decrease in MAPE. This condition was also proven in various scenarios of the number of training sets, and MCMTD significantly influenced the decrease in error rate. SMOTE-RkNN hybrid sampling focuses on oversampling by calculating nearest neighbors to the resulting samples. SMOTE-RkNN has been tested using thirty datasets with binary classes. Experimental results showed that SMOTE-RkNN was superior in ten datasets using the classification and regression tree (CART) classifier, eighteen datasets using the linear discriminant analysis (LDA) classifier, and twenty datasets using the Gaussian naive bayes (GNB) classifier. This condition also shows that SMOTE-RkNN has a significant influence on classification performance.

HybridDA is a hybrid approach that combines data level and algorithm to address imbalanced data. The data level is used to address imbalanced data, and the algorithm level is used to optimize using grid search. The grid search optimization is performed through cost parameters, gamma, and kernel support vector machines (SVM). Experimental results using the Portuguese marketing campaign dataset obtained optimal results compared to the baseline. CSMOUTE hybrid sampling combines synthetic majority undersampling technique (SMUTE) and synthetic majority oversampling technique (SMOTE). This approach has been tested using the knowledge extraction based on evolutionary learning (KEEL) repository dataset with numerical data types and binary classes. Experimental results showed that CSMOUTE obtained optimal results using multi-layer perceptron (MLP) and support vector machine (SVM) classifiers, while using the logistic regression (LR) classifier was not optimal. Finally, MC-CCR is an approach focused on addressing imbalanced data for multi-class datasets. MC-CCR has been experimented with using nineteen numerical datasets. Experimental results showed that MC-CCR was superior in eleven datasets or could be categorized as having a significant influence on improving classification performance.

Previous research experiments using various hybrid approaches such as MCMTD, SMOTE-RkNN, HybridDA, CSMOUTE, and MC-CCR, have shown several factors that can impact classification performance in addressing imbalanced data. These factors include the type of data in the dataset, the class (binary or multi-class), and the classification algorithm. Consequently, the proposed hybrid sampling approaches have different characteristics while solving problems. It is important to highlight that all hybrid sampling approaches have both advantages and disadvantages, depending on the problem-solving design. However, conducting more experiments on various types of datasets can result in a more stable and robust approach, ultimately influencing perspectives on the hybrid approach to be used.

Table 4. Proposed imbalanced data methods

Method	Level	Class	Dataset	Avg. Acc
MCMTD [18]	Data	Binary	Number	94,47%
SMOTE-RkNN [19]	Data	Binary	Number	95%
HybridDA [20]	Data	Binary	Combine	96.73%
CSMOUTE [21]	Data	Binary	Number	95%
MC-CCR [22]	Data	Multi	Number	97,12%
ROS-NCL	Data	Multi	Text	97,94%

In this study, ROS-NCL is proposed as an alternative solution to imbalanced data problems. It has been tested using text data with multi-class and compared to existing approaches such as MCMTD, SMOTE-RkNN, HybridDA, CSMOUTE, and MC-CCR. Although this proposed approach may produce different performance when applied and tested using other data, it has been compared to other hybrid approaches such as SMOTE-Tomek and SMOTE-ENN, which have shown that ROS-NCL produces better and more stable performance. Additionally, ROS-NCL has been tested using three types of classification model, namely LSTM, CNN, and CNN-LSTM, and has been proven to obtain superior performance.

4. Conclusion and future work

The aim of this study is to address the issue of imbalanced multi-class text data classification tasks through the development of a novel and efficient preprocessing method. To achieve this goal, a hybrid sampling ROS-NCL algorithm is proposed that integrates both oversampling and undersampling methods. The approach was designed to improve the resolution of class imbalance in multi-class datasets, which can be a major challenge for machine or deep learning algorithms. The effectiveness of the proposed hybrid sampling solution using a curated smart city dataset was evaluation, and the experimentation confirmed that the approach significantly improves the performance of imbalanced multi-class data classification tasks.

In particular, the research indicates that the proposed methodology surpasses other conventional approaches for managing imbalanced data. The results of this study showcase the efficacy of the

preprocessing technique in tackling the issue of imbalanced multi-class text data classification. By integrating both oversampling and undersampling methods, a more resilient resolution to class imbalance challenges in multi-class datasets is offered. The technique is readily applicable to various text data classification tasks and has the potential to enhance the precision and dependability of machine or deep learning algorithms.

This study has made a significant contribution to the field of machine or deep learning by proposing a novel and efficient preprocessing method to handle imbalanced multi-class text data classification tasks. However, it is important to note that the research has some limitations that require further investigation. Firstly, the proposed hybrid sampling algorithm focused exclusively on addressing imbalanced text data. Future studies should explore the effectiveness of our method on other types of imbalanced datasets, such as those containing images or numerical data. Secondly, the classification architectures employed, namely LSTM, CNN, and CNN-LSTM, are still conventional. Further development is needed to produce superior and stable performance, particularly in the context of complex, real-world datasets. Finally, the proposed method was not evaluated on other datasets, beyond the smart city dataset already curated. Future study should entail performance evaluation across various datasets, to better understand the generalizability of this study's approach and its potential limitations.

Notwithstanding the aforementioned limitations, the study has laid the groundwork for future investigations concerning the treatment of class imbalance in multi-class datasets, specifically within the realm of text data. The proposed hybrid sampling algorithm exhibits the potential to enhance the performance of machine or deep learning algorithms for various imbalanced datasets, and the authors encourage further inquiry in this domain.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

Conceptualization, UE and AFR; methodology, UE and AW; software, UE; validation, AFR. and AW; formal analysis, UE; investigation, UE and AW; resources, UE; data curation, UE; writing original draft preparation, UE; writing review and editing, UE; visualization, UE; supervision, AFR. and AW; project administration, UE.

Acknowledgments

We would like to express our sincere gratitude to Universitas Diponegoro, particularly the Doctoral Program of Information Systems, for providing us with the necessary facilities and resources to conduct this study. We also would like to express our sincere gratitude to Universitas Bina Darma as the funding agency that supported the publication of this study. Finally, we extend our thanks to the editors, reviewers, and staff of the publishing journal for their careful attention and guidance throughout the peer review and publication process. Their expertise and professionalism have been invaluable in bringing this work to fruition.

References

[1] E. Strelcenia and S. Prakoonwit, "Improving Classification Performance in Credit Card Fraud Detection by Using New Data Augmentation", *AI*, Vol 4, No.1, pp. 172–198, 2023.

[2] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance", *Journal of Big Data*, Vol.6, No.1, pp. 2-54, 2019.

[3] C. Fan, X. Li, Y. Zhao, and J. Wang, "Quantitative assessments on advanced data synthesis strategies for enhancing imbalanced AHU fault diagnosis performance", *Energy and Buildings*, Vol. 252, No. December, pp. 1-16, 2021.

[4] M. A. Alfhaid and M. Abdullah, "Classification of Imbalanced Data Stream: Techniques and Challenges", *Transactions on Machine Learning and Artificial Intelligence*, Vol.9, No.2, pp. 36–52, 2021.

[5] S. Das, S. S. Mullick, and I. Zelinka, "On Supervised Class-Imbalanced Learning: An Updated Perspective and Some Key Challenges", *IEEE Transactions on Artificial Intelligence*, Vol. 3, No. 6, pp. 973–993, 2022.

[6] Y. Guo, J. Feng, B. Jiao, N. Cui, S. Yang, and Z. Yu, "A dual evolutionary bagging for class imbalance learning", *Expert Systems with Applications*, Vol. 206, No. May, p. 117843, 2022.

[7] V. S. Spelmen and R. Porkodi, "A Review on Handling Imbalanced Data", In: *Proc. of International Conf. on Current Trends towards Converging Technologies*, Coimbatore, India, pp. 1–11, 2018.

[8] N. N. Nguyen and A. T. Duong, "Comparison of two main approaches for handling imbalanced data in churn prediction problem", *Journal of Advances in Information Technology*, Vol. 12, No. 1, pp. 29–35, 2021.

[9] W. Zhang, X. Li, X. D. Jia, H. Ma, Z. Luo, and X. Li, "Machinery fault diagnosis with imbalanced data using deep generative adversarial networks", *Measurement: Journal of the International Measurement Confederation*, Vol. 152, No. May, p. 107377, 2020.

[10] K. Upadhyay, P. Kaur, S. Prasad, and L. Vidyapeeth, "State of the Art on Data level methods to address Class Imbalance Problem in Binary Classification", *GIS Science Journal*, Vol. 8, No.3, pp. 875-903, 2021.

[11] F. S. Gharehchopogh, "An Improved Harris Hawks Optimization Algorithm with Multi-strategy for Community Detection in Social Network", *Journal of Bionic Engineering*, Vol. 19, No. 4, pp. 1177-1202, 2022.

[12] F. S. Gharehchopogh, "An Improved Tunicate Swarm Algorithm with Best-random Mutation Strategy for Global Optimization Problems", *Journal of Bionic Engineering*, Vol. 19, No. 4, pp. 1177-1202, 2022.

[13] F. S. Gharehchopogh, A. Ucan, T. Ibrikci, B. Arasteh, and G. Isik, "Slime Mould Algorithm: A Comprehensive Survey of Its Variants and Applications", *Archives of Computational Methods in Engineering*, Vol. 20 No. 1. pp. 158-183, 2023.

[14] F. S. Gharehchopogh, M. Namazi, L. Ebrahimi, and B. Abdollahzadeh, "Advances in Sparrow Search Algorithm: A Comprehensive Survey", *Archives of Computational Methods in Engineering*, Vol. 30, No. 1, pp. 427–455, 2023.

[15] F. S. Gharehchopogh, "Advances in Tree Seed Algorithm: A Comprehensive Survey", *Archives of Computational Methods in Engineering*, Vol. 29, No. 5, pp. 3281–3304, 2022.

[16] F. S. Gharehchopogh, "Quantum-inspired metaheuristic algorithms: comprehensive survey and classification", *Artificial Intelligence Review*, Vol. 55, No.7, pp. 1-65, 2022.

[17] H. Mohammadzadeh and F. S. Gharehchopogh, "Feature Selection with Binary Symbiotic Organisms Search Algorithm for Email Spam Detection", *International Journal of Information Technology & Decision Making*, Vol. 20, No. 1, pp. 469–515, 2021.

[18] X. Yu, Y. He, Y. Xu, and Q. Zhu, "A Mega-Trend-Diffusion and Monte Carlo based virtual sample generation method for small sample size problem", *Journal of Physics: Conf. Series*, Vol. 1325, No. 1, pp. 1-6, 2019.

[19] A. Zhang, H. Yu, Z. Huan, X. Yang, S. Zheng, and S. Gao, "SMOTE-RkNN: A hybrid re-sampling method based on SMOTE and reverse k-nearest neighbors", *Information Sciences*, Vol. 595, No. May, pp. 70–88, 2022.

[20] M. M. A. Rifaie and H. A. Alhakbani, "Handling Class Imbalance In Direct Marketing Dataset Using A Hybrid Data And Algorithmic Level Solutions", In: *Proc. of SAI Computing Conference, London*, United Kingdom, pp. 446-451, 2016.

[21] M. Koziarski, "CSMOUTE: Combined Synthetic Oversampling and Undersampling Technique for Imbalanced Data Classification", In: *Proc. of International Joint Conf. on Neural Networks, Shenzhen*, China, pp. 1-8, 2021.

[22] M. Koziarski, M. Woźniak, and B. Krawczyk, "Combined Cleaning and Resampling Algorithm for Multi-Class Imbalanced Data With Label Noise", *Knowledge-Based Systems*, Vol. 204, No. September, p. 106223, 2020.

[23] B. Krawczyk, M. Koziarski, and M. Wozniak, "Radial-based oversampling for multiclass imbalanced data classification", *IEEE Transactions on Neural Networks and Learning Systems*, Vol.31, No.8, pp. 2818–2831, 2020.

[24] S. Sharma, C. Bellinger, B. Krawczyk, O. Zaiane, and N. Japkowicz, "Synthetic Oversampling with the Majority Class: A New Perspective on Handling Extreme Imbalance", In: *Proc. of IEEE International Conf. on Data Mining (ICDM)*, Singapore, pp. 447–456, 2018.

[25] K. Agustianto and P. Destarianto, "Imbalance Data Handling using Neighborhood Cleaning Rule (NCL) Sampling Method for Precision Student Modeling", In: *Proc. International Conf. on Computer Science, Information Technology, and Electrical Engineering*, Jember, Indonesia, pp. 86–89, 2019

[26] M. Khushi, K. Shaukat, T. M. Alam, I. A. Hameed, S. Uddin, S. Luo, X. Yang, and M. C. Reyes, "A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data", *IEEE Access*, Vol. 9, No. August, pp. 109960–109975, 2021.

[27] H. Faris, "Neighborhood Cleaning Rules and Particle Swarm Optimization for Predicting Customer Churn Behavior in Telecom Industry", *International Journal of Advanced Science and Technology*, Vol. 68, No. 2, pp. 11–22, 2014.

[28] U. Ependi, A. F. Rochim, and A. Wibowo, "Smart City Assessment for Sustainable City Development on Smart Governance: A Systematic Literature Review", In: *Proc. of International Conf. on Decision Aid Sciences and Applications, Chiangrai*, Thailand, pp. 1088–1097, 2022.

[29] M. Pira, "A novel taxonomy of smart sustainable city indicators", *Humanities and Social Sciences*

Communications, Vol. 8, No. 1, pp. 1-8, 2021.

[30] A. Akande, P. Cabral, P. Gomes, and S. Casteleyn, "The Lisbon ranking for smart sustainable cities in Europe", *Sustainable Cities and Society*, Vol. 44, No. August, pp. 475–487, 2019.

[31] P. Berrone and J. E. Ricart, "IESE Cities in Motion Index 2018", *Barcelona, IESE Business School's*, 2018.

[32] ITU-T, *Key performance indicators for smart sustainable cities to assess the achievement of sustainable development goals*, 2022. [Online]. Available: https://www.itu.int/rec/T-REC-Y.4903

[33] A. J. Benites and A. F. Simões, "Assessing the urban sustainable development strategy: An application of a smart city services sustainability taxonomy", *Ecological Indicators*, Vol. 127, No. August, p. 107734, 2021.

[34] B. Cohen, *Smart city index master indicators survey. Smart cities council*, 2022. [Online]. Available: http://smartcitiescouncil.com

[35] G. Koca, O. Egilmez, and O. Akcakaya, "Evaluation of the smart city: Applying the dematel technique", *Telematics and Informatics*, Vol. 62, No. June, p. 101625, 2021.

[36] F. Purnomo, Meyliana, and H. Prabowo, "Smart city indicators: A systematic literature review", *Journal of Telecommunication, Electronic and Computer Engineering*, Vol. 8, No. 3, pp. 161–164, 2016.

[37] R. M. Kimiya and S. A. Torabi, "Ranking cities based on their smartness level using MADM methods", *Sustainable Cities and Society*, Vol. 72, No. May, p. 103030, 2021.

[38] L. Shen, Z. Huang, S. Wai, S. Liao, and Y. Lou, "A holistic evaluation of smart city performance in the context of China", *Journal of Cleaner Production*, Vol. 200, No. November, pp. 667–679, 2018.

[39] V. F. Anez, G. Velazquez, and F. P. Prada, "Smart City Projects Assessment Matrix: Connecting Challenges and Actions in the Mediterranean Region", *Journal of Urban Technology*, Vol. 27, No. 4, pp. 1–25, 2018.

[40] R. A. Sharif and S. Pokharel, "Smart City Dimensions and Associated Risks: Review of literature", *Sustainable Cities and Society*, Vol. 77, No. February, p. 103542, 2022.

[41] T. Yigitcanlar, K. Degirmenci, L. Butler, and K. C. Desouza, "What are the key factors affecting smart city transformation readiness ? Evidence from Australian cities", *Cities*, Vol. 120, No. August, p. 103434.

[42] S. Gutman and P. Vorontsova, "Issues of Development of Smart Transport Assessment Indicators", In: *Proc of International Scientific Conf. on Digital Transformation on Manufacturing, Infrastructure and Service*, Sofia, Bulgaria, 2020.

[43] C. Patrão, P. Moura, and A. T. D. Almeida, "Review of Smart City Assessment Tools", *Smart Cities*, Vol. 3, No. 4, pp. 1117–1132, 2020.

[44] Y. Zhou, P. Yi, W. Li, and C. Gong, "Assessment of city sustainability from the perspective of multi-source data-driven", *Sustainable Cities and Society*, Vol. 70, No. January, p. 102918, 2021.

[45] J. J. E. Macrohon, C. N. Villavicencio, X. A. Inbaraj, and J. Jeng, "A Semi-Supervised Approach to Sentiment Analysis of Tweets during the 2022 Philippine Presidential Election", *Information*, Vol. 13, No. 484, pp. 1–14, 2022.

[46] A. D. Widiantoro, A. Wibowo, and B. Harnadi, "User Sentiment Analysis in the Fintech OVO Review Based on the Lexicon Method", In: *Proc. of Sixth International Conf. on Informatics and Computing*, Jakarta, Indonesia, pp. 1–4, 2021.

[47] W. López, J. Merlino, and P. R. Bocca, "Learning semantic information from Internet Domain Names using word embeddings", *Engineering Applications of Artificial Intelligence*, Vol. 94, No. September, p. 103823, 2020.

[48] A. Abdi, S. M. Shamsuddin, S. Hasan, and J. Piran, "Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion", *Information Processing & Management*, Vol. 56, No. 4, pp. 1245–1259, 2019.

[49] I. K. Sastrawan, I. P. A. Bayupati, and D. M. S. Arsa, "Detection of fake news using deep learning CNN–RNN based methods", *ICT Express*, Vol. 8, No. 3, pp. 396–408, 2022.

[50] M. A. Diniz, "Statistical methods for validation of predictive models", *Journal of Nuclear Cardiology*, Vol. 29, No. 6, pp. 3248–3255, 2022.

[51] T. E. Tallo and A. Musdholifah, "The Implementation of Genetic Algorithm in Smote (Synthetic Minority Oversampling Technique) for Handling Imbalanced Dataset Problem", In: *Proc. of International Conf. on Science and Technology*, pp. 1–4, 2018.

[52] C. Rana, N. Chitre, B. Poyekar, and P. Bide, "Stroke Prediction Using Smote-Tomek and Neural Network", In: *Proc. of International Conference on Computing Communication and Networking Technologies*, Kharagpur, India, pp.

1-5, 2021.

[53] G. Menardi and N. Torelli, "Training and assessing classification rules with imbalanced data", *Data Mining and Knowledge Discovery*, Vol. 28, No. 1, pp. 92-122, 2014.

[54] M. T. Vásquez, O. C. Bosquez, B. H. Ocaña, and J. H. Torruco, "Classification of guillain-barre syndrome subtypes using sampling techniques with binary approach", *Symmetry*, Vol. 12, No.3, pp. 1–27, 2020.

[55] K. Agustianto and P. Destarianto, "Imbalance Data Handling using Neighborhood Cleaning Rule (NCL) Sampling Method for Precision Student Modeling", In: *Porc. of International Conf. on Computer Science, Information Technology, and Electrical Engineering*, Jember, Indonesia, pp. 86–89, 2019.

[56] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey", *Information*, Vol. 10, No. 4, pp. 1–68, 2019.

[57] G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification", *Neurocomputing*, Vol. 337, No. April, pp. 325–338, 2019

[58] A. Bhavani and B. S. Kumar, "A Review of State Art of Text Classification Algorithms", In: *Proc. of International Conf. on Computing Methodologies and Communication*, Erode, India, pp. 1484–1490, 2021.

[59] A. Hikmah, S. Adi, and M. Sulistiyono, "The Best Parameter Tuning on RNN Layers for Indonesian Text Classification", In: *Proc. of International Seminar on Research of Information Technology and Intelligent Systems*, Yogyakarta, Indonesia, pp. 94–99, 2020.

[60] M. O. Ibrohim, E. Sazany, and I. Budi, "Identify abusive and offensive language in indonesian twitter using deep learning approach", *Journal of Physics: Conference Series*, Vol. 1196, No. 1, pp. 267-278, 2019.

[61] A. F. Hidayatullah, A. M. Hakim, and A. A. Sembada, "Adult content classification on Indonesian tweets using LSTM neural network", In: *Proc. of International Conf. on Advanced Computer Science and information Systems*, Bali, Indonesia, pp. 235–240, 2019.

[62] R. Saputra, A. Waworuntu, and A. Rusli, "Classification of Indonesian News using LSTM-RNN Method", In: *Proc. of International Conf. on New Media Studies*, Tangerang, Indonesia, pp. 72–77, 2021.

[63] S. Wu, Y. Liu, Z. Zou, and T. H. Weng, "S_I_LSTM: stock price prediction based on multiple data sources and sentiment analysis", *Connection Science*, Vol. 34, No. 1, pp. 44–62, 2022.

[64] X. Chen, P. Cong, and S. Lv, "A Long-Text Classification Method of Chinese News Based on BERT and CNN", *IEEE Access*, Vol. 10, No. March, pp. 34046–34057, 2022.

[65] Imamah, U. L. Yuhana, A. Djunaidy, and M. H. Purnomo, "Development of Text Classification Based on Difficulty Level in Adaptive Learning System using Convolutional Neural Network", In: *Proc. of International Electronics Symposium, Surabaya*, Indonesia, pp. 238–243, 2021.

[66] E. M. Dharma, F. L. Gaol, H. L. H. S. Warnars, and B. Soewito, "The Accuracy Comparison Among Word2Vec, Glove, and Fasttext Towards Convolution Neural Network (CNN) Text Classification", *Journal of Theoretical and Applied Information Technology*, Vol. 100, No. 2, pp. 349–359, 2022.

[67] D. Alsaleh and S. L. M. Sainte, "Arabic Text Classification Using Convolutional Neural Network and Genetic Algorithms", *IEEE Access*, Vol. 9, No. June, pp. 91670–91685, 2021.

[68] S. Saadah, K. M. Auditama, A. A. Fattahila, F. I. Amorokhman, A. Aditsania, and A. A. Rohmawati, "Implementation of BERT, IndoBERT, and CNN-LSTM in Classifying Public Opinion about COVID-19 Vaccine in Indonesia", *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, Vol. 6, No. 4, pp. 648–655, 2022.

[69] F. Meng, W. Cheng, and J. Wang, "Semi-supervised software defect prediction model based on tri-training", *KSII Transactions on Internet and Information Systems*, Vol. 15, No. 11, pp. 4028–4042, 2021.

[70] L. Bencke, C. Cechinel, and R. Munoz, "Automated classification of social network messages into Smart Cities dimensions", *Future Generation Computer Systems*, Vol. 109, No. August, pp. 218–237, 2020.