2022 5th International Conference on Networking, Information Systems and Security: Envisage Intelligent Systems in 5g//6G-based Interconnected Digital Worlds (NISS) | 978-1-6654-5363-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/NISS55057.2022.10085156

Synonym-based Text Generation in Restructuring Imbalanced Dataset for Deep Learning Models

Febi Siti Sutria Ningsih*‡, Purnomo Husnul Khotimah†§, Andria Arisal†, Andri Fachrur Rozie†, Devi Munandar†, Dianadewi Riswantini†, Ekasari Nugraheni†, Wiwin Suwarningsih†, Dian Kurniasari*

*Faculty of Mathematics and Natural Sciences
University of Lampung
Bandar Lampung, Indonesia
‡sitifebi20@gmail.com

†Research Center for Informatics
National Research and Innovation Agency
Bandung, Indonesia
§purn005@brin.go.id

***Abstract*—One of which machine learning data processing problems is imbalanced classes. Imbalanced classes could potentially cause bias towards the majority classes due to the nature of machine learning algorithms that presume that the object cardinality in classes is around similar number. Oversampling or generating new objects in minority class are common approaches for balancing the dataset. In text oversampling method, semantic meaning loses often occur when deep learning algorithms are used. We propose synonym-based text generation for restructuring the imbalanced COVID-19 online-news dataset. Three deep learning models (MLP, CNN, and LSTM) using TF/IDF and word embedding (WE) feature are tested with the original and balanced dataset. The results indicate that the balance condition of the dataset and the use of text representative features affect the performance of the deep learning model. Using balanced data and deep learning models with WE greatly affect the classification significantly higher performances as high as 4%, 5%, and 6% in accuracy, precision, recall, and f1-score, respectively.**

***Index Terms*—text generation, imbalanced dataset, deep learning**

I. Introduction

Machine learning is an inferring technique from a dataset that uses a mathematical approach to create a model that reflects patterns in the data. Machine learning has two minimum goals. That is, predicting the future and acquiring knowledge [1]. There are many problems with machine learning data processing, one of which is related to data class imbalances.

There is a data class imbalance because there is a main class (minority) and another class that contains a significantly higher number of data (majority). For example, imbalance classes in the case of identifying COVID-19 infection event using online news. The proportion of articles that actually report the infection event is quite smaller than articles that report information surrounding the COVID-19 pandemic, such as health tips, fund raise for the impacted group, etc. The benefit of using online news to detect infectious disease event is essential for quick detection, specially when it is an international event. With the hierarchical system of world health reporting system, the infection event happened in other countries may be too slow to be identified.

Machine learning algorithm is canonical and assume that objects occurences in each classes is roughly similar [2]. Therefore, it is difficult to classify data into minority classes if the data class sharing is not balanced. To minimize the number of errors, the learning algorithm relies only on most of the classification classes [3].

To balance imbalanced dataset requires data preprocessing so that the models developed in data mining and machine learning algorithms are efficient. Resampling is a method of dealing with imbalanced problems by iterating over partial data and distributing balanced data through random iterations [4].

There have been multiple studies that investigating in the area of data resampling in imbalanced data [5]–[8]. Statistical Learning with imbalanced data [5] conclude that the performance of different classification algorithms does not necessarily indicate a change to the different sampling algorithms. There are several classifiers that have high sensitivity to sampling algorithms such as SVM, Random Forest, K-Nearest Neighbors, Decision Tree and mentions the oversampling algorithm shows better performance than undersampling. The same thing was stated by Garcia [8] who carried out research to see the effect of imbalance and classifier performance by performing several resampling strategies for imbalanced datasets [6] comparison of performance isn't it using general measurements such as overall accuracy, precision, and recall, but with F-Measure, G-Mean, and AUC with results showing that using the oversampling method produces better performance, but there is no oversampling method which out performs other methods. Oversampling can cause overfitting. Therefore, Chawla [7] proposes a solution to handle overfitting by the resampling method is oversampling using the SMOTE (Synthetic Minority Oversampling Technique) method by utilizing Nearest Neighbors and the amount of oversampling for numerical

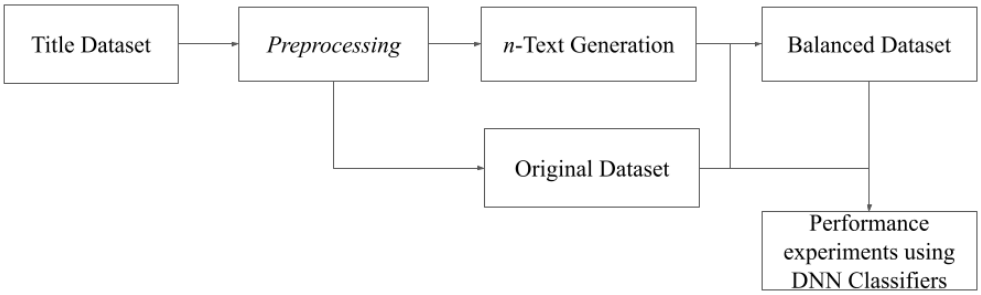

Fig. 1. Framework of this study

data approaches.

However, Synthetic data generation and oversampling techniques like SMOTE and ADASYN can overcome this problem for statistics data, but these methods anguish a substantial overfitting and noise of model. While the technique has proven useful for generation of numerical data and synthetic images using GANs, the effectiveness of the proposed approach for textual data, which can keep grammatical structure, context, and semantic information, has not been evaluated [9]. Additionally, in the case of Indonesian language, it suffers resource scarcity problem that is limited availability of data collection and Natural Language Processing (NLP) libraries [10].

In this study, we propose an n-text generation for oversampling. n-phrases from the original text are chosen randomly to be replaced with their synonyms and generate n-new texts. We use the Kateglo API to generate synonyms for the selected phrases. Three common deep learning models (MLP, CNN, and LSTM) are used and their performances are compared using the original Indonesian COVID-19 online news dataset and the balanced dataset.

II. Related Studies

Data imbalance often occurs in real world cases. For example, user results queries commonly returns a small number of results that match with user's parameters [11]. Another example is when classify news articles received from several news agencies of interest to specific users [12]. Imbalanced classification also an important problem in other that text domain, such as for medical diagnosis, deceit detection and many other problems. In this study, we focus on text domain.

The distribution of imbalanced data for each classes is oblique because some classes appear more frequently compared to other classes. This condition causes the classifier will show a bias to the majority class and for some extreme cases, may overlook the minority class at all. This condition is not preferred because quite often the minority class data is the class that represent the interest concept [9].

Several approached are used in dealing imbalance dataset, namely: data-level approach, algorithm-level approach and and hybrid approach [2]. In the case of algorithm and hybrid-level approaches, it involves with modifying avalaible classifiers or libraries to alleviate the bias towards majority class. Due to NLP libraries for Indonesian corpus is still lacking, we concentrate in data-level approach in this study.

For data level approach, data restructuring is done by modifying the training data to match the standard learning algorithm [2]. The restructuring is done by balancing the data is undersampling the method that removing samples from majority class or make synthetic samples by generating for minority class (oversampling). The problems with undersampling is that it often leads to removal of important samples [2]. Additionally, in cases where the minority number is small, the classifier will only have a small dataset to learn from. Hence, oversampling is a potential approach in our study case.

In text generation, oversampling is to resampling the data by creating new text. Some studies has been conducted by using deep learning to create new samples [9]. However, unlike numeric and images data, text data may risk of loosing its semantic and contextual information. The resulting text would be in bad text structure and grammar thus the meaning is loss. Therefore, we propose a text generation method using n-phrase synonym replacement. n-phrase from a text will be replaced by its synonym in order to create similar new text. Hence, the new text will retain its meaning. The detail of the proposed methods is explained in Methodology section.

III. Methodology

The research framework to be carried out is illustrated in the in Fig. 1. The raw data is the COVID-19 online news title. The title is then preprocessed and entered the n-Text Generation function. The generated text is added with the original dataset to produce the balanced dataset. Both original dataset and balanced dataset will be tested by comparing the performance of deep learning models (DNN Classifiers). A complete methodology used in our framework is detailed in the following explanation. In addition, we explain the dictionary library (Kateglo API) and the feature used for the deep learning classifiers, namely TF/IDF and word embedding.

A. Title Dataset

The COVID-19 Indonesian online news data is obtained from crawling online news from various Indonesian news portals with the keyword "covid" [13]. The dataset consists of 4 columns, namely publication time, news portal, news title, and types of news event. Types of news event are classified into

TABLE I
THE DISTRIBUTION OF THE ONLINE NEWS DATASET

Online News Portal	Non-event	Event	Ratio
Antara	7,103	3,090	70:30
Detik	2,394	1,609	60:40
Kompas	948	797	54:40
Kumparan	1,318	450	75:25
Merdeka	3,825	2,592	60:40
Republika	6,927	5,204	57:43
Tempo	4,149	2,926	59:41

two types, which is event (1) and non-events (0). Event type is the class of interest that is articles that report COVID-19 infection event. As for non-event type is a class of articles that report information surrounding COVID-19 event. The event type is manually labelled by three respondents and a voting mechanism is used to assign the final label.

The collected data is as much as 16,884 data from seven different Indonesian news portals such as "Antara", "Detik", "Kompas", "Kumparan", "Merdeka", "Republika", and "Tempo". Minority class data (event news) which has been separated is as much as 4,549 data. The rest is the majority class (non-event) that is as much as 12,295 data. Table I shows the distribution of the online news dataset.

B. Preprocessing

Because the data used is in the form of unstructured text, which stores a lot of valuable information but have a little common structural framework, a preprocessing stage is required. In natural language processing, the preprocessing cost more than 60% of the workload [14]. The following issues may caused the high load: the use of inappropriate grammar, spelling errors, local dialects and semantic ambiguity that increase the complexity of data processing and analysis. However, we used trusted online news portal. Hence, such issues should be minimal. Nonetheless, we conducted several basic preprocess as follows:

- case folding will be carried out by changing the size of the word letters to lowercase
- text cleaning to remove non-alphabet characters
- exclude the stopwords. Indonesian stopwords are used for sorting out words according to the requirements Indonesian language [15].

C. n-Text Generation

Text generation is done by changing some of the words in the news title into synonyms that will be taken from Kateglo [16] as Indonesian dictionary. Fig. 2 shows the procedure of the n-text generation. n-phrases will be randomly pick from the original text. In each iteration of n, the synonym of the picked phrase is retrieved from Kateglo and the phrase in the original text will be replaced by the similar word.

To do the n-text generation, we first need to count the n. n is the parameter that define the number of how many new text be generated. n number is decided by the following formula : $n = ceil(\Delta(\Sigma_{majorityclass} - \Sigma_{minorityclass})/\Sigma_{minorityclass})$.

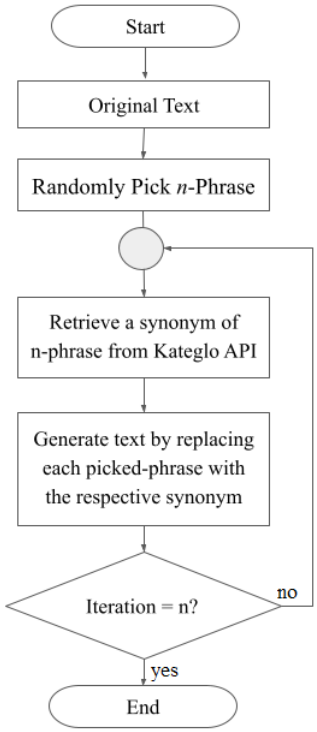

Fig. 2. Flow Chart n-Text Generation

Using the formula, we attained $n = 2$. This number means that from each event news title (4,945 data) will be generated 2 new titles. As the result, the generated text consist of 9,049 new data. Since, we only require 7,746 data to balance the data, the generated text is randomly selected. In the end, the balanced data will be in the size of 24,630 data.

D. Kateglo API

Kateglo API is a web service that provides definitions, synonyms, antonyms, formal entries, and glossaries related to words and phrases. The name kateglo comes from the abbreviation of the service elements ka(mus), te(saurus), and glo(sarium). Kateglo is an open source application based on PHP and MySQL data, consisting of 72,253 dictionary entries 191,000 glossary entries, 2,012 proverb entries, and several Indonesian dictionary entries with 3,423 abbreviation and acronym entries. Kateglo can be used manually by entering the required keywords and can also be used for applications that provide output in JSON or XML format [16].

For this research, we used the Kateglo to find several similar of each word in the original text and them will be generated. We will find the similarities of a particular word by passing parameters to the URL. Below is an example of the URL:
`http://kateglo.com/api.php?format=json&phrase=[word]`

```
{'kateglo':{'actual_phrase':None,
'all_relation' : [{'rel_uid':'224363'
'root_phrase':'biaya',
'related_phrase':'anggaran',
'rel_type':'s',
'updated':null,
'updater':'TESAURUS',
'rel_type_name':'Sinonim',
'lex_class':'n'},
```


Fig. 3. Kateglo API

TABLE II
NEWS TITLE BEFORE AND AFTER THE WORDS REPLACED BY ITS SYNONYM

Original title	Generated title
pasien reaktif hasil rapid test covid kabur dari rs bahteramas sultra (Eng. the patient reactive result of covid rapid test escaped from the bahteramas hospital north sulawesi)	- *pasien reaktif akibat rapid test covid kabur dari rs bahteramas sultra* (Eng. a patient reactive due to of covid rapid test escaped from the bahteramas hospital north sulawesi) - *pasien reaktif hasil rapid test covid menghilang dari rs bahteramas sultra* (Eng. a reactive patient result of covid rapid test disappear from the bahteramas hospital north sulawesi)
seluruh pasien positif covid di belitung sembuh pariwisata pun segera pulih (Eng. all positive covid patients in Belitung recover tourism will soon recover)	- *segenap pasien positif covid di belitung sembuh pariwisata pun segera pulih* (Eng. all positive covid patients in Belitung recover tourism will soon recover) - *seluruh pasien positif covid di belitung sembuh pariwisata pun lekas pulih* (Eng. all positive covid patients in Belitung recover tourism will quickly recover

For example, when the word ”biaya” is chosen, a response from Kateglo server in the format of JSON data will be received [17]. The result from from Kateglo is shown at Fig 3. The JSON data will be parsed and used particularly when the value of JSON dictionary data in `['kateglo']['all_relation'] ['rel_type']` is equal to `'s'`. Table II shows examples of the news title before and after the words replaced by its synonym. The colored text (red and blue) shows its respective synonym.

E. TF-IDF and Word Embedding

Term Frequency-Inverse Document Frequency (TF-IDF) and Word Embedding (WE) are two common techniques for text representation. They are used to convert terms into numerical values. TF-IDF uses statistical method to determine the importance of terms within document collection [18]. The idea is that each words has different level of importance in a document collections. The value is depend on term frequency (TF), inverse document frequency (IDF), and document length. TF is the cardinality of frequent words occurences in the document, meanwhile IDF is a weighted sum documents on the corpus and the documents number with appearance of a word [19].

On the other hand, Word Embedding (WE) uses language modelling and feature learning from specified document corpus to map words into vector of real numbers. It utilizes various neural network techniques [20].

F. Classifier Model

In the experiment we use three main deep learning model, as the following:

- MLP (multi layer perceptron) is a neural network architecture contains several layers comprising of an input layer, one or more hidden layers, and an output layer. Nonlinear activation functions are used in every node in the hidden layers and output layer. Layers construct a feed-forward structure in which the output of one layer becomes the input of the next layer [21], [22].
- CNN (convolutional neural network) is an MLP regularization that used to keep away overfitting by adding a convolution layer as a regularizer [23]. At the end of the process, the classification function is provided by adding two types of layers contained in the CNN, namely a fully connected layer and an output layer which is similar to MLP.
- LSTM (long short term memory) [24] is an extension of neural network architecture with the ability to study dependencies between data in a dataset. LSTM has units where the main points are cells that control the flow of information from input gates to output gates or forget gates for unimportant information.

Fig 4 shows the model for the respective DNN models.

IV. RESULT AND DISCUSSION

In this study, we experimented with MLP, CNN, and LSTM (deep learning models) using the datasets that had been generated, namely the original data (imbalanced) and the new data generated (balanced). The feature used to represent the text is TF-IDF and keras word embedding layer. Those models were evaluated using k-fold cross validation. To keep away the small amounts of training and testing data of each fold, we use k = 5 in place of 10, epoch = 100, where the optimizer is adam. In addition, we use early stopping based on loss value [25].

The performance metrics that will be used are listed below [26]:

- Accuracy is a measurement of performance between the predicted data processing and the actual total data proccessed.
- Precision is a performance measurement which is the proportion of data, which is correctly predicted positive to the total predicted positive from the processed data.

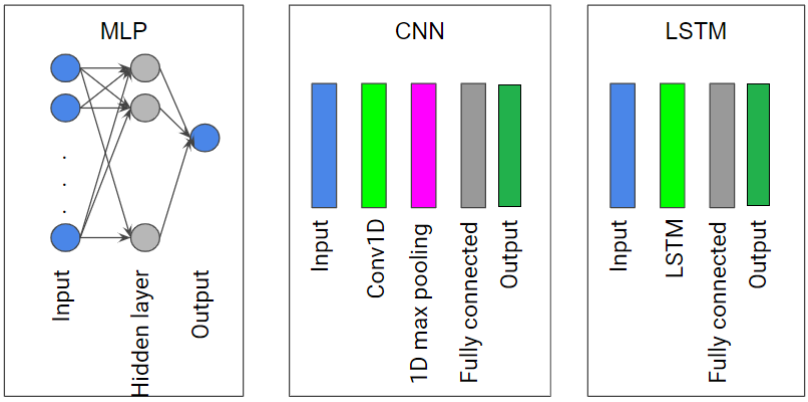

Fig. 4. The DNN diagram

TABLE III
METRICS COMPARISON ACCURACY DNN-TFIDF

	Original Dataset				Balanced Dataset			
Classifier	Acc.	Prec.	Recall	F1-score	Acc.	Prec.	Recall	F1-score
MLP	72.96	64.27	50.50	43.70	63.57	63.61	63.59	63.56
CNN	73.69	72.05	52.19	47.07	66.65	67.20	66.64	66.37
LSTM	79.55	67.94	65.12	63.40	80.74	81.51	80.75	80.63

TABLE IV
METRICS COMPARISON ACCURACY DNN-WE

	Original Dataset				Balanced Dataset			
Classifier	Acc.	Prec.	Recall	F1-score	Acc.	Prec.	Recall	F1-score
MLP	91.28	89.32	88.35	88.81	95.01	95.05	95.02	95.01
CNN	91.98	90.25	89.22	89.70	94.62	94.70	94.64	94.62
LSTM	91.32	89.24	88.63	88.91	93.49	93.54	93.49	93.49

- Recall is the proportion of positive predicted positive data to all actual processed data in the true class.
- F1-Score is a performance measurement which is average comparison of precision and recall values.

The results are shown in Table IV and Table V.

Based on Table III, we can see that the accuracy value of the DNN-TFIDF using the original data (imbalanced) is stable at the average 70%. However, for the recall value and f1-score on MLP, CNN has a low value of consecutive average 50% and 40%. This does not apply to LSTM which has a stable value precision, recall, and F1-score at 60%. Meanwhile for the balanced data, MLP and CNN shows a decrease of performance in the value of accuracy (60%). However, we can observe that all four metrics for both MLP and CNN values show a stable score at 63% and 66-67%, respectively. These scores are in contrast with the the same DNN models using original data. Using original data, both MLP and CNN's precision, recall and F1-score are decreasing. Additionally, the LSTM achieves the highest performance and it was stable at a value of 80% for each metric.

As for the DNN-WE, both balance and imbalanced data attained a significantly higher performance compare with DNN-TFIDF as shown in Table IV. The result implied that for text classification, word embedding more suitable choice than TF-IDF. Further, using imbalanced data, the DNN-WE models increases the accuracy value to 91%; precision are in the range of 89-90%; recall and f1-score are stable at range 88-89%. Additionally, all DNN-WE models using balanced data increase significantly in their performances by 2-4%. In accuracy, DNN-WE increase by 2-4%. The same phenomenon can be observed in DNN-WE using balanced data that is all four metrics' score are more stable compared to the one that using the original dataset. LSTM has the best performance using balanced dataset with accuracy in 93.49%, precision in 93.54%, 93.49%, and F1-score in 93.49%.

For additional discussion, we would like to address a few issues of the generated text. We found that a small portion of the generated text could create a possible noise as listed in Table V. Limiting the words that can be replaced by its synonym can be a prospective solution for future improvements.

V. CONCLUSION

From this experiment, we found that Deep Learning is better to strive for using balanced data. Although we haven't been able to do experiment using other datasets, it can be seen

TABLE V
POSSIBLE NOISE IN THE GENERATED TEXT

Issue item	Original	Generated text
Replacement by uncommon word	*kasus covid bertambah hingga mei dan imbauan tak berkerumun* Eng. covid cases increase until May and appeal not to crowd	*kasus covid babar hingga mei dan imbauan tak berkerumun* Eng. covid cases babar until May and appeal not to crowd
Replacement of location name	*covid di as sudah mencapai kasus kematian* Eng. covid in the us has reached a death case	*covid di aksis sudah mencapai kasus kematian* Eng. covid in axis has reached death cases

that balanced data can produce higher and stable performance values compared to imbalanced data. Our experiment shows that the accuracy performance increases as high as 4%. In addition, using word embedding as the text feature for DNN models, the performance value is better. So it can be concluded that the data imbalance and the addition of word embedding greatly affect the classification testing of deep learning models such as MLP, CNN, and LSTM. For future work, we are interested to deeper explore in synonym based text generation.

REFERENCES

[1] J. W. G. Putra, "Pengenalan konsep pembelajaran mesin dan deep learning," *Tokyo. Jepang*, 2019.

[2] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.

[3] S. Mutmainah, "Penanganan imbalance data pada klasifikasi kemungkinan penyakit stroke," *Jurnal Sains, Nalar, dan Aplikasi Teknologi Informasi*, vol. 1, no. 1, 2021.

[4] M. F. Akbar, I. Kurniawan, and A. Fauzi, "Mengatasi imbalanced class pada software defect prediction menggunakan two-step clustering-based undersampling dan bagging tehcnique," *Jurnal Informatika*, vol. 6, no. 1, pp. 107–113, 2019.

[5] A. Shipitsyn, "Statistical learning with imbalanced data," 2017.

[6] H. He, Y. Bai, E. Garcia, and S. A. Li, "adaptive synthetic sampling approach for imbalanced learning. ieee international joint conference on neural networks," in *2008 (IEEE World Congress On Computational Intelligence)*, 2008.

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[8] V. García, J. S. Sánchez, and R. A. Mollineda, "On the effectiveness of preprocessing methods when dealing with different levels of class imbalance," *Knowledge-Based Systems*, vol. 25, no. 1, pp. 13–21, 2012.

[9] S. Shaikh, S. M. Daudpota, A. S. Imran, and Z. Kastrati, "Towards improved classification accuracy on highly imbalanced text dataset using deep neural language models," *Applied Sciences*, vol. 11, no. 2, p. 869, 2021.

[10] B. Wilie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, S. Soleman, R. Mahendra, P. Fung, S. Bahar *et al.*, "Indonlu: Benchmark and resources for evaluating indonesian natural language understanding," *arXiv preprint arXiv:2009.05387*, 2020.

[11] A. Y.-c. Liu, "The effect of oversampling and undersampling on classifying imbalanced text datasets," Ph.D. dissertation, Citeseer, 2004.

[12] A. Sun, E.-P. Lim, and Y. Liu, "On strategies for imbalanced text classification using svm: A comparative study," *Decision Support Systems*, vol. 48, no. 1, pp. 191–201, 2009.

[13] E. Nugraheni, P. H. Khotimah, A. Arisal, A. F. Rozie, D. Riswantini, and A. Purwarianti, "Classifying aggravation status of covid-19 event from short-text using cnn," in *2020 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*. IEEE, 2020, pp. 240–245.

[14] W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang, "Data processing and text mining technologies on electronic medical records: a review," *Journal of healthcare engineering*, vol. 2018, 2018.

[15] D. Munandar, A. F. Rozie, and A. Arisal, "A multi domains short message sentiment classification using hybrid neural network architecture," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 4, pp. 2181–2191, 2021.

[16] Z. Saniyah, "Normalisasi mikroteks berbentuk singkatan pada teks twitter berbahasa indonesia menggunakan algoritma longest common subsequences," 2019.

[17] M. A. Fauzi, N. Firmansyah, T. Afirianto *et al.*, "Improving sentiment analysis of short informal indonesian product reviews using synonym based feature expansion," 2018.

[18] A. Rajaraman and J. D. Ullman, *Mining of massive datasets*. Cambridge University Press, 2011.

[19] P. M. R. C. Dinatha and N. A. Rakhmawati, "Komparasi term weighting dan word embedding pada klasifikasi tweet pemerintah daerah," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi— Vol*, vol. 9, no. 2, 2020.

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[21] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Representation sby Back-Propagating Errors," *Nature*, vol. 323, pp. 533–536, 1986, doi: https://doi.org/10.1038/323533a0.

[22] S. K. Pal and S. Mitra, "Multilayer perceptron, fuzzy sets, classification," *IEEE Transactions on Neural Networks*, vol. 3, no. 5, pp. 683–697, 1992, doi: https://doi.org/10.1109/72.159058.

[23] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," *Proceedings of the 25th International Conference on Machine Learning*, pp. 160–167, 2008, doi: https://doi.org/10.1145/1390156.1390177.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: https://doi.org/10.1145/1390156.1390177.

[25] P. H. Khotimah, A. F. Rozie, E. Nugraheni, A. Arisal, W. Suwarningsih, and A. Purwarianti, "Deep learning for dengue fever event detection using online news," in *2020 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*. IEEE, 2020, pp. 261–266.

[26] U. Salamah, "Perbandingan metode pembelajaran mesin berbasis parametrik dan non-parametrik untuk klasifikasi diabetic retinopathy imagery," *JSAI (Journal Scientific and Applied Informatics)*, vol. 4, no. 2, pp. 193–198, 2021.