A Robustly Optimized BERT using Random Oversampling for Analyzing Imbalanced Stock News Sentiment Data

2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE) | 979-8-3503-2095-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCoSITE57641.2023.10127725

Salsabila Mazya Permataning Tyas
Department of Informatics Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
salsa25mazya@gmail.com

Riyanarto Sarno
Department of Informatics Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

Agus Tri Haryono
Department of Informatics Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
agustharyono24@gmail.com

Kelly Rossa Sungkono
Department of Informatics Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
kelly@its.ac.id

Abstract— **Stock news is one of the information sources thatcan used to monitor stock prices. The information from stock news usually contains positive and negative sentiments that canaffect stock prices. Therefore, sentiment analysis is needed to process the sentiment of stock news. The stock news dataset is taken from Kaggle. From these data, there is an imbalanced class between positive and negative sentiment. This research proposed a method to solve the imbalance dataset with random oversampling which worked by randomly replicating several minority classes. This research presents several scenarios of pre-processing text with different stages, intending to get high accuracy. The classification method used in this paper is a robustly optimized Bidirectional Transformer Encoder Representation (RoBERTa). Besides that, this paper also compared with baseline of Machine Learning (ML) such as Multinomial Naïve Bayes, Bernoulli Naïve Bayes, Support Vector Machine, Random Forest Classifier, Logistic Regressionand used two different text representation such as TF-IDF and Word2Vec. The best result in this research is obtained using RoBERTa method with the fourth scenario of pre-processing text, in which the stage of pre-processing in this scenario only removing hashtag, without removing punctuation, removing thenumber, converting number, stop word removal, and lemmatization. The performance result is 0.85 precision, 0,84 recall, 0,84 F1-score, and 86% for accuracy result.**

Keywords-pre-processing text, random oversampling, robustly optimized BERT, sentiment analysis, stock news.

I. INTRODUCTION

In these last few years, stock investment has become a worldwide trend. The Stock market is a place for investors to seek short-term (trading) and long-term investment profit. Choosing a good sector or company will take much work because of the stock market's volatile natural environment [1]. Therefore, sentiment analysis is needed on information sources about stock price changes in the stock market.

One of the most crucial information sources is financial news. Previous research revealed that news could affect stockprices in the stock market [2]. The stock market price depends on information news, events, and company announcements rather than past and present stock prices. However, using computational algorithms to automatically analyze text using Natural Language Processing (NLP) is one possible solution for proper analyzing a significant news volume. Sentiment analysis is a task in NLP that attempts to identify the feeling of texts, whether positive, negative, or neutral. In this research, the sentiment analysis applied to financial news could increase the quality of financial agent's decisions.

Sentiment analysis has a vital and essential task in text mining, Natural Language Processing (NLP), and information retrieval, namely text pre-processing [3], [4]. It is employed in the preparation of unstructured data for knowledge extraction. Previous study [5] stated that preprocessing of the data is a very important step as it decides the efficiency of the other steps down in line for sentiment analysis, they carried out several pre-processing stages; namely stemming words, remove the hashtag, quotes, remove question marks, remove the https links using regular expression. One of our research objectives is to improve that not all datasets case canbe carried out in the Pre-processing Text stage to get high- accuracy results. Therefore, in our experiment several preprocessing text scenarios were carried out to get higher accuracy results, this is one of the differences from previous studies.We did some stages of pre-processing text. There are removing punctuation, removing hashtags, stop word removal, and lemmatization.

In previous studies [6] BERT for stock market sentiment analysis. According to this study, the method used is BERT with a pre-trained BERT-base model. These cutting-edge models were developed by optimizing the pre-trained BERT model with an additional output layer, obtaining an accuracyof 82.5% with a total of 582 articles of data from seven sources websites with different compositions. An imbalanced dataset is a manageable problem because it can indicate of low accuracy. The purpose of our research is not only to notice the preprocessing stage of the text but also the composition of the dataset. To handle this problem, we used oversampling echnique. Oversampling is a technique to overcome the problem of unbalanced data by processing the minority classso that the proportion in the sample is

greater than the original proportion. There are many oversampling techniques to solvethe imbalanced dataset problem. The type of oversampling that we used is random oversampling. Random oversampling generates new samples by randomly sampling with replacement of the currently available sample.

In addition to previous studies, the method used was BERT which used transformer-based pre-trained models. A development model called RoBERTa is currently being developed to overcome the BERT model's undertraining. RoBERTa is a robust optimization BERT model with a significant difference in the modified hyperparameters and relates to how differently the model is trained [7]. In our research, the RoBERTa model will be combined without stop word removal, lemmatization, and removing punctuation in preprocessing text which can increase accuracy by 0.4 if compared with using several-stage preprocessing text.

Therefore, the objective of this work is to experimentallyassess the imbalanced data by a random oversampling technique using RoBERTa method in stock news sentiment analysis to get the high accuracy. The primary contributions of this effort thus far are given below.

- Resolve the composition of the amount of imbalance data by using the random oversampling technique
- Conducted several experimental scenarios in the preprocessing text stage for the unstructured Stock News dataset with the aim of finding the best accuracy
- Experimental evaluation comparing several Machine learning and deep learning models especially RoBERTa methods.

Section 1 (Introduction) has discussed the sentiment analysis problem with an imbalanced dataset and scenario of preprocessing text to get the best accuracy. Section 2 (Related theories) explains theoretical approaches relevant to the writer's research. Then, section 3 (Methodology) describes the methodology of the proposed method. We focused on thestages of this research. Section 4 (Result) reports and discusses the experiment result. Furthermore, section 5 (Conclusion) explains concludes the experiment discussion.

II. RELATED WORKS

A. Sentiment Analysis

Sentiment analysis is the process for analyze an opinion, evaluation, and assessment related to the topic, individual, and organization. Besides that, sentiment analysis is one branch of Natural Language Processing (NLP) used to recognize and process the subjective data from text [8].

B. Pre-processing Text

Pre-processing text usually consists of stages: stop word removal, case folding, lemmatization, tokenization, removing numbers and punctuation. This stage aims to reduce the noiseof datasets and get high accuracy. In this research we try to did all stages in pre-processing text. We analyzed the cause of thelow value of accuracy. So, we did several experiments to increase accuracy value. There are four scenarios of pre- processing text in this research.

C. Random Oversampling

In this case, there are imbalanced data sets. Imbalanced datasets perform excellently in one data class but not inanother class. A previous study [9] stated two categories ofthe previous algorithm for solving an imbalanced problem: methods based on classifier modification and pre-processing.To tackle this problem in this research, we used random oversampling. Random oversampling shown in Fig. 1 worked by randomly replicating several minority classes. In theprevious study [10], agree that random oversampling canincrease the accuracy value by considering imbalance dataset, because it make an exact copy of the minority sample class.

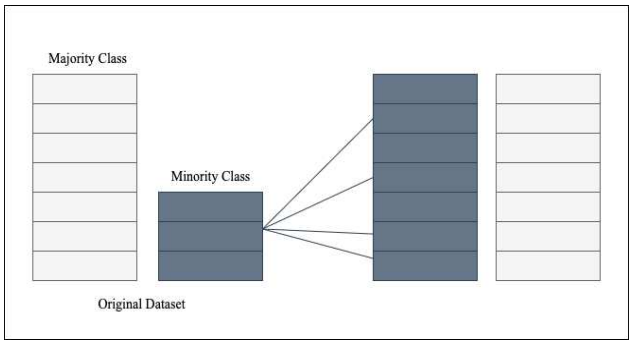

Fig.1 Random Oversampling

D. A Robustly Optimized BERT

Bidirectional Encoder Representation from Transformer (BERT) is a transformer model of bidirectional encoders [11]. This model capture the context between the layer and can be represented to combine left and proper context [12].

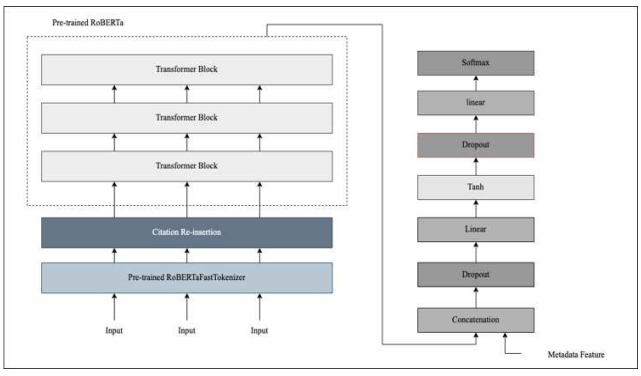

Fig. 2 Architecture RoBERTa

Wenxiong et al. [13] stated that BERT's variants better performed better than BERT. RoBERTa is a robustly optimized BERT, which modified some hyperparameters in BERT model. The main difference is how differently the model is trained. RoBERTa is a transformer model that is pre-trained on a large corpus of English data in a self-supervised manner. The RoBERTa model was pre-trained on the reassembly of five data sets which together these

datasets weigh 160GB of text, there are: BookCorpus, English Wikipedia, CC-News, OpenWebText, Stories. There are twomodels of RoBERTa based on parameters, Roberta-base and Roberta-large. The difference between these models is the parameter size. Roberta-base uses 12-layer and 768-hidden layers, while Roberta-large uses 24-layer and 1024-hidden layers. Architecture of this model shown in Fig. 2.

E. Evaluation

In this section, according to Imam Ghozali et al. [14], we used accuracy, precision, F1-score, and recall to measure theperformance evaluation. And then, we represented the performance of RoBERTa model evaluation with a confusion matrix. The performance of data testing with ground truth value represented by true positive (TP), false positive (FP), true negative (TN) and false negative (FN) determines the recall and precision calculating procedure.

. Precision (P), as shown in Equation (1), is the exactness value between the ground truth information and the response decided by the system.

$$Precision = \frac{TP}{(TP+FP)} \quad (1)$$

Recall (R) is the success rate to regain information by the system, as explained in Equation (2) [15].

$$Recall = \frac{TP}{(TP+FN)} \quad (2)$$

Accuracy is defined as the level of closeness between the predicted value and the actual value, as shown in Equation (3).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

F1-Score is an illustration of comparison of the weighted average precision and recall, as shown in Equation (4).

$$F1 - Score = 2\,x\,\frac{Precision\ x\ Recall}{Precision+Recall} \quad (4)$$

III. METHODOLOGY

A. Dataset

Gathered stock news dataset from multiple twitter handles regarding financial news. This is a public dataset, we took from Kaggle. There are 5.764 data, which divides into two sentiments. Negative sentiment is represented by (-1), which means this tweet has the potential to affect stock price going down, while positive sentiment represented with (1), this tweet can potentially affect stock price rise.

There is an imbalanced composition of the dataset, which count of negative sentiment is 2.106, and the count of positive sentiment is 3.685.

B. Pre-processing Text

Procedure:

Start
Step #01: Take a tweet from dataset in previous stage
Step #02: Use the Python Natural Language Toolkit (NLTK) to doing pre-processing text
Step #03: Do Remove punctuation → Remove hashtag → Stop word removal → Lemmatization.
Step #04: Save the result for next stage.
End

Fig. 3 Pseudocode Pre-processing Text

At this stage, we did four scenarios to process the dataset in pre-processing text. Fig. 3 shows the pseudocode of general pre-processing text, and we also remove tweets whose text length is less than three words. The example of each process represented Table I. The scenario of the pre- processing text that we do is in Table II. First, in stages scenario 1 of pre-processing text, there are removing punctuation, removing hashtag, stop word removal and lemmatization. Then, in scenario two, we add remove the number from the scenario. We try to remove numbers because after we analyze the stock market dataset, some datacontain numbers, so we want to know whether removing numbers can increase or decrease the accuracy value. In the third scenario, we add the process of the first scenario by converting the number and symbol percent "&" to text. For example, the number "21" changed to "twenty one" and changed the symbol "%" to "percent". Last scenario, we did not remove punctuation, remove the number, convert the number, stop word removal, or lemmatization. In this stage, we are only doing remove hashtags.

TABLE I. Examples pre-processing text

Pre-processing	Raw Data	Data Cleaning
Remove Punctuation	User: AIG (American International Group) Option Traders bet on 4% down moved by next Friday #stock	User AIG American International Group Option Traders bet on 4 down moved by next Friday #stock
Remove Hashtag	User AIG American International Group Option Traders bet on 4 down moved by next Friday #stock	User AIG American International Group Option Traders bet on 4 down moved by next Friday stock
Stop word Removal	User AIG American International Group Option Traders bet on 4 down moved by next Friday stock	User AIG American International Group Option Traders 4 down moved next Friday stock
Lemmatization	User AIG American International Group Option Traders 4 down moved next Friday stock	User AIG American International Group Option Traders 4 down move next Friday stock
Remove number	User AIG American International Group Option Traders 4 down move next Friday stock	User AIG American International Group Option Traders down move next Friday stock
Convert number to text	User AIG American International Group Option Traders 4 down move next Friday stock	User AIG American International Group Option Traders four down move next Friday stock

TABLE II. Scenario of Pre-processing Text

Scenario	Stage of Pre-processing Text
Scenario 1	(Remove punctuation, Remove hashtag, Stop word removal, Lemmatization)
Scenario 2	(Remove punctuation, Remove hashtag, Stop word removal, Lemmatization) + Remove number
Scenario 3	(Remove punctuation, Remove hashtag, Stop word removal, Lemmatization) + Convert number and symbol "%" to text.
Scenario 4	(Remove hashtag)

C. Balancing Data

After the pre-processing text, the count of a dataset is reduced. Because there is some tweet that the length less than three word was deleted. Fig. 4 shows that the composition of a dataset is imbalanced. We did random oversampling to solve this imbalanced problem. It works by randomly replicating several minority classes.

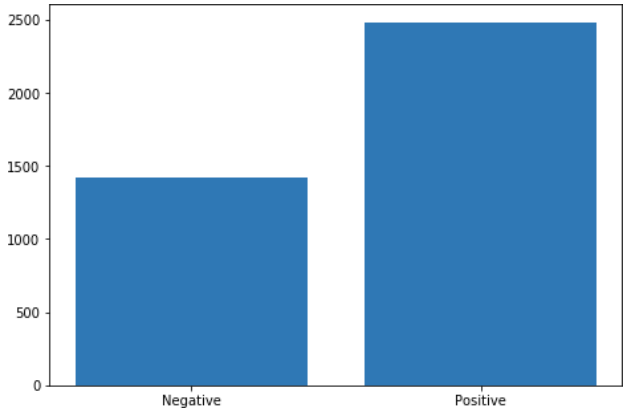

Fig. 4 Imbalanced Dataset

D. Classification

In this study, the primary method we used is RoBERTa, which is an optimizer from the BERT model. We conducted experiments comparing baseline methods based on machine learning. We used two types text representation: TF-IDF and word embedding Word2vec. We used five baseline methods of machine learning that we used, there are Multinomial Naïve Bayes, Bernoulli Naïve Bayes, Support Vector Machine, Random Forest Classifier, and Logistic Regression. Each baseline method will be processed with two text representations such as TF-IDF and Word2Vec.

E. Evaluation

By comparing results, the sentiment classification performance of each method was evaluated. Performance is measured using precision, recall, accuracy, and F1-score.

IV. Result

A. Result Balancing Data

This study has two major limitations. First, this study used imbalanced dataset. The One of our contributions is to solve the problem of the imbalanced dataset. To deal with this problem, we use random oversampling to reduce the class imbalanced of the dataset and make the data distribution stabilized [16]. The count of the original dataset afterthe pre-processing text is 2.349 for sentiment positive and 1.363 for sentiment negative. After doing random oversampling, based on Table III, we get the final proportion of the dataset, which is 2.349 for sentiment positive and 2.349 for sentiment negative. Furthermore, random oversampling technique makes the data have an equivalent value and the data class has become an average value.

TABLE III. Count of Dataset After Random Oversampling

Sentiment	Total Dataset
Negative	2349
Positive	2349

B. Classification Result

We used deep learning and machine learning based approaches. The algorithms of machine learning such as Multinomial Naïve Bayes, Bernoulli Naïve Bayes, Support Vector Machine, Random Forest Classifier, and Logistic regression and two deep learning algorithms such as BERT and RoBERTa. The result of each method's classification, BERT and RoBERTa represented in Table IV.

In measuring model performance, the selected Roberta model successfully recognizes the dataset with its corpus butneeds to be more successful in certain sentence forms. The example "BIOF wants 4.90's comin!!!" was misclassified. The second limitation of this research, because the data used is sentiment about stocks, so that the data contains a lot of numbers. We need to analyze these numbers whether they have an effect or not, by carrying out pre-processing text scenarios. Therefore, we compared four scenarios of pre-processing text the RoBERTa method: Scenario 1 (Remove punctuation, remove hashtag, stop word removal, lemmatization), Scenario 2 (Remove punctuation, remove hashtag, stop word removal, lemmatization) + remove the number, Scenario 3 (Remove punctuation, remove hashtag, stop word removal, lemmatization) + convert number and symbol "%" to text, and the last scenario, Scenario 4 it's only removed hashtag.

According to Table IV, the hightest obtained by RoBERTa-base with scenario 4 of pre-processing text was better than another scenario. The result of pre-processing textwithout removing punctuation, removing number, stop wordremoval, and lemmatization gave higher accuracy. The best performance of RoBERTa-base with scenario 4 obtained was86% for accuracy value

We compared our deep learning results, especially RoBERTa, with the machine learning method, according to Table V. Experimenting based on the machine learning method, we used general pre-processing text, which means we used Scenario 1 of pre-processing text. There are two model text representations such as TF-IDF and Word2Vec toprocess each method such as Multinomial Naïve Bayes, Bernoulli Naïve Bayes, Support Vector Machine, Random Forest Classifier, Logistic Regression. Between the two model text representation, accuracy with model text representation TF-IDF got high score than Word2Vec,

and there is Support Vector Machine with 78% accuracy. Nevertheless, if we compared it with the main experiment RoBERTa, the experimental results continue to improve when scenario 4 applied which only remove hashtag. This advantage yields the best accuracy is 86%.

TABLE IV. The Result of Deep Learning Algorithm

Model	Pre-trained	Scenario of pre-processing	Precision	Recall	F1-score	Accuracy
BERT	Bert-base	Scenario 1	0.78	0.78	0.78	79%
	Bert-base	Scenario 1	0.80	0.79	0.79	81%
RoBERTa	RoBERTa-base	Scenario 1	0.81	0.81	0.81	82%
	RoBERTa-base	Scenario 2	0.32	0.50	0.39	63%
	RoBERTa-base	Scenario 3	0.72	0.72	0.72	74%
	RoBERTa-base	Scenario 4	0.85	0.84	0.84	86%
	RoBERTa-large	Scenario 1	0.32	0.50	0.39	63%

TABLE V. The Result of Machine Learning Algorithm

Model	Pre-trained	Precision	Recall	F1-score	Accuracy
TF-IDF	Multinomial Naïve Bayes	0.73	0.74	0.73	74%
	Bernoulli Naïve Bayes	0.72	0.73	0.73	74%
	Support Vector Machine	0.77	0.73	0.74	78%
	Random Forest Classifier	0.75	0.74	0.74	77%
	Logistic Regression	0.75	0.76	0.75	77%
Word2Vec	Multinomial Naïve Bayes	0.61	0.61	0.61	64%
	Support Vector Machine	0.75	0.71	0.72	76%
	Random Forest Classifier	0.69	0.62	0.62	70%
	Logistic Regression	0.69	0.66	0.67	71%

V. CONCLUSIONS

This research solves the imbalance case on the sentiment of the stock news dataset with a random oversampling technique. We compared four scenarios of pre-processing text to get high accuracy. It shows that removing hashtag resulting the highest accuracy. The comparison of the classifier methods, it shows that a robust optimized BERT get the highest result. For future work, we suggest using anothertechnique for handling imbalanced datasets, like Synthetic Minority Oversampling Technique (SMOTE) as an ensemble oversampling and other transformer models, like XLM RoBERTa which has a larger corpus.

ACKNOWLEDGMENT

This research was funded by Lembaga Pengelola Dana under Riset Inovatif-Produktif (RISPRO) Program; The Indonesian Ministry of Education and Culture under Penelitian Terapan Unggulan Perguruan Tinggi (PTUPT) Program; the Indonesian Ministry of Education, Culture, Research and Technology through the Matching Fund (MF) Kedaireka Program; and Institut Teknologi Sepuluh Nopember (ITS) through the Article Processing Charge (APC) Incentive Program.

REFERENCES

[1] A. T. Haryono, R. Sarno, and R. Abdullah, 'Aspect-Based Sentiment Analysis of Financial Headlines and Microblogs Using Semantic Similarity and Bidirectional Long Short-Term Memory', *Int. J. Intell. Eng. Syst.*, 2022.

[2] Y. Ren, F. Liao, and Y. Gong, 'Impact of News on the Trend of Stock Price Change: an Analysis based on the Deep Bidirectiona LSTM Model', *Procedia Comput. Sci.*, vol. 174, pp. 128–140, 2020, doi: 10.1016/j.procs.2020.06.068.

[3] D. S. Vijayarani and J. Ilamathi, 'Preprocessing Techniques for Text Mining - An Overview', vol. 5.

[4] M. A. Palomino and F. Aider, 'Evaluating the Effectiveness of Text Pre-Processing in Sentiment Analysis', *Appl. Sci.*, vol. 12, no. 17, p. 8765, Aug. 2022, doi: 10.3390/app12178765.

[5] J. Abraham, D. Higdon, J. Nelson, and J. Ibarra, 'Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis', vol. 1, no. 3, 2018.

[6] M. G. Sousa, K. Sakiyama, L. de S. Rodrigues, P. H. Moraes, E. R. Fernandes, and E. T. Matsubara, 'BERT for Stock Market Sentiment Analysis', in *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, Portland, OR, USA, Nov. 2019, pp. 1597–1601. doi: 10.1109/ICTAI.2019.00231.

[7] G. R. Narayanaswamy, 'Exploiting BERT and RoBERTa to Improve Performance for Aspect Based Sentiment Analysis', 2021, doi: 10.21427/3W9N-WE77.

[8] B. S. Rintyarna, R. Sarno, and C. Fatichah, 'Evaluating the performance of sentence level features and domain sensitive features of product reviews on supervised sentiment analysis tasks', *J. Big Data*, vol. 6, no. 1, p. 84, Dec. 2019, doi: 10.1186/s40537-019-0246-8.

[9] H. Ali *et al.*, 'A review on data preprocessing methods for class imbalance problem', *Int. J. Eng.*.

[10] M. Hayaty, S. Muthmainah, and S. M. Ghufran, 'Random and Synthetic Over-Sampling Approach to Resolve Data Imbalance in Classification', *Int. J. Artif. Intell. Res.*, vol. 4, no. 2, p. 86, Jan. 2021, doi: 10.29099/ijair.v4i2.152.

[11] H. Xu, B. Liu, L. Shu, and P. S. Yu, 'BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis'. arXiv, May 03, 2019. Accessed: Jan. 08, 2023. [Online]. Available: http://arxiv.org/abs/1904.02232

[12] Z. Gao, A. Feng, X. Song, and X. Wu, 'Target-Dependent Sentiment Classification With BERT', *IEEE Access*, vol. 7, pp. 154290–154299, 2019, doi: 10.1109/ACCESS.2019.2946594.

[13] W. Liao, B. Zeng, X. Yin, and P. Wei, 'An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa', *Appl. Intell.*, vol. 51, no. 6, pp. 3522–3533, Jun. 2021, doi: 10.1007/s10489-020-01964-1.

[14] I. Ghozali, K. R. Sungkono, R. Sarno, and R. Abdullah, 'Synonym based feature expansion for Indonesian hate speech detection', *Int. J. Electr. Comput. Eng. IJECE*, vol. 13, no. 1, p. 1105, Feb. 2023, doi: 10.11591/ijece.v13i1.pp1105-1112.

[15] Institut Teknologi Sepuluh Nopember, R. Priyantina, R. Sarno, and Institut Teknologi Sepuluh Nopember, 'Sentiment Analysis of Hotel Reviews Using Latent Dirichlet Allocation, Semantic Similarity and LSTM', *Int. J. Intell. Eng. Syst.*, vol. 12, no. 4, pp. 142–155, Aug. 2019, doi: 10.22266/ijies2019.0831.14.

[16] A. Mukherjee, S. Mukhopadhyay, P. K. Panigrahi, and S. Goswami, 'Utilization of Oversampling for multiclass sentiment analysis on Amazon Review Dataset', in *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)*, Morioka, Japan, Oct. 2019, pp. 1–6. doi: 10.1109/ICAwST.2019.8923260.