ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 16, No. 4, pp. 1225-1239, July 2025
DOI: https://doi.org/10.17507/jltr.1604.17

Sentence Classification Using Machine Learning and Word Embedding: An Innovation in Indonesian Language Learning

Sri Kusuma Winahyu
Research Center for Language, Literature, and Community, National Research and Innovation Agency (BRIN), Indonesia

Fawwaz Zaini Ahmad
Bank Rakyat Indonesia, Indonesia

Achril Zalmansyah*
Research Center for Language, Literature, and Community, National Research and Innovation Agency (BRIN), Indonesia;
Doctoral Program in Education, FKIP-Lampung University, Indonesia

Exti Budihastuti
Research Center for Language, Literature, and Community, National Research and Innovation Agency (BRIN), Indonesia

Pradicta Nurhuda
Research Center for Language, Literature, and Community, National Research and Innovation Agency (BRIN), Indonesia

Fairul Zabadi
Research Center for Language, Literature, and Community, National Research and Innovation Agency (BRIN), Indonesia

Zainal Abidin
Research Center for Preservation Language and Literature, National Research and Innovation Agency (BRIN), Indonesia

Suyadi
Research Center for Language, Literature, and Community, National Research and Innovation Agency (BRIN), Indonesia

Sri Yono
Research Center for Language, Literature, and Community, National Research and Innovation Agency (BRIN), Indonesia

Evi Maha Kastri
Research Center for Preservation Language and Literature, National Research and Innovation Agency (BRIN), Indonesia

***Abstract*—In applied linguistics, writing assessment examines language learning. There are various genres in writing, but the evaluation always includes a syntactic component or sentence structure. This research focuses on classifying sentence structure in the Indonesian language using the Random Forest Classifier algorithm on five different experiment models, which are trained using different vectorization techniques, including bag of word (BoW), hashing, Term Frequency-Inverse Document Frequency (TF-IDF), CBoW, and skipgram vectorizers. The results showed that the accuracy of the models varied significantly, with the highest accuracy of 76% achieved by the model trained using the CBOW vectorizer. The model trained using the BoW vectorizer and skipgram vectorizer had the lowest accuracies of 65%. These results suggest that different vectorization techniques significantly impact the accuracy of the model and the CBoW vectorization technique is the most**

* Corresponding Author. Email: zzalmansa@gmail.com; ORCID iD: https://orcid.org/0000-0002-3883-5463

effective. While the skipgram was trained using the dataset itself before being used to vectorize the dataset, but it did not show a significant improvement in accuracy. Classifying sentence structures with various models is important and may continue to support the syntactic assessment of computer-based Indonesian language writing skills.

***Index Terms*—Bag of Word, Continuous Bag of Word, sentence structure, syntactic assessment, Term Frequency-Inverse Document Frequency, vectorization techniques**

I. Introduction

Language is used as a means of communication only by humans. This is due to the human ability to create a symbol or name something. This is an ability that is not possessed by animals or plants. It also functions as a system of language symbols in the culture of a nation (Rabiah, 2018; Tektigul et al., 2023; Zalmansyah, 2017). Language may contribute to science, socio-culture, and society (Hariyanto et al., 2023; Nardiati et al., 2023; Nursugiharti et al., 2024; Zalmansyah et al., 2023). In this modern era, language cannot be separated from science and technology, so it involves the use of technology as part of language development.

Technological developments and advances touch various aspects of life, including in the field of language (Alimyar & Lakshmi G, 2021). The COVID-19 pandemic, which broke out some time ago, has created the need for an online learning system. After the pandemic is relatively over, the "positive" side is that online learning has become an alternative to the reality of education. In the online learning process, the assessment aspect is one of the parts.

This research examines the detection of Indonesian sentence structures using machine learning and word embeddings. The importance of creating a model or an Artificial Intelligence that can detect a sentence structure is to assist in the automatic assessment of the result of writing texts. Writing texts are produced by students in the language learning process at school. In Indonesia, writing is trained and tested in the form of text production. In the 2013 Curriculum (revised) text-based learning has been applied to Indonesian Language subject. Writing skills as one of the four language skills (listening-reading-writing-speaking) have especially begun to be taught at the elementary school level in text form with various types of text genres (De Smedt et al., 2016; Fatonah & Wiradharma, 2018; Hoyos Pipicano, 2024; Jagaiah et al., 2020; Jiang et al., 2022; Khair & Misnawati, 2022; Kim & Zagata, 2024; Nordin et al., 2022; Philippakos et al., 2023; Renza et al., 2022; Rodríguez-Gonzalo & Abad-Beltrán, 2023; Setiawan et al., 2019; Sinaga et al., 2023; Sun et al., 2022; Zalmansyah, 2017; Zalmansyah, 2018).

In all text genres, there must be elements of words, phrases, clauses, sentences, and paragraphs. If the assessment is carried out on text, one of the elements of text scoring is syntax scoring which if conducted automatically, requires a model or script that can detect every sentence structure in a text. However, this research is limited only to the detection system at the sentence level and does not extend to paragraphs or scoring systems in the form of scores.

The components of the Indonesian writing assessment generally consist of content, organization, syntax, vocabulary, and mechanics (Abeywickrama & Brown, 2010). The detection of mechanical components and vocabulary in written Indonesian texts has been carried out in an Indonesian spelling editing application, called SIPEBI, developed by Mayani (Ramliyana et al., 2022; Utami, 2022; Winahyu, 2024). This application is connected to the Indonesian language dictionary, called KBBI (Moeljadi et al., 2016) for matching with the correct vocabulary elements. Meanwhile, Ratna also developed a web-based automatic essay scoring system, SIMPLE-O. This system works using latent semantics analysis (LSA) which compares the text with the words chosen as references (Ahmad & Laroche, 2023; Kim et al., 2020; Ratna et al., 2015; Valdez et al., 2018; Venugopal et al., 2023).

The syntactic component in this case is related to the use of sentences with the proper sentence structure or function syntax. Therefore, in this study, the data used is in the form of sentences with a standard structure. So, if there is a sentence that uses a non-standard structure, the machine will not read it, so the sentence is categorized as a wrong sentence.

With the advent of Machine Learning and Big Data, machine learning is constantly applied to solve various problems. The idea behind big data and the implementation of machine learning is to gather large enough data, and let a machine learning algorithm analyze those data, then form a model from patterns that are found in those analyzed data. Those patterns that are found by the algorithms are normally invisible to the naked human eye. This is the main reason why we implemented machine learning to create a model that could detect sentence structure in the Indonesian language. Since in the Indonesian language, a particular word could act as more than one entity depending on the position in a sentence, thus we could not implement Named Entity Recognition and POS tagging.

II. Literature Review

Algorithms based on transformational grammar have been used for a long time in detecting sentences, especially English sentences (Arooj et al., 2024; Campbell et al., 2012; Friedman, 1969, 1971; Leacock et al., 2011; Lee et al., 2014; Wortmann & Stouffs, 2018). In addition, in the English language, efforts to detect sentences as support for automatic writing assessments have indeed been started since the 1960s, although they then vacated and reappeared in 1970, initiated by Ellis Page and a team at the University of Connecticut who developed an automated essay scoring engine (Shermis et al., 2010).

Research on the detection of Indonesian sentence structure with machine learning is still rare. Some of them are research conducted by Gunawan, Mudafiq, and Sulastra. Identify and analyze sentence structure in the Indonesian language, even though not using a machine learning algorithm *per se*. They used LALR (Look Ahead Left Right) parser and POS (Part of Speech) Tag to identify Bahasa Indonesia sentence structure (Gunawan et al., 2019). The research focused on *kalimat tunggal* or otherwise known as a single sentence. First sentences are gathered, then preprocessing, and later processing, then finally post-processing. Preprocessing includes a set of processes that cleans the sentences from punctuation and tokenization, which outputs a tokenized sentence with the label of each token. Meanwhile, the processing stage took those labeled tokens to the LALR parser. There are no training processes in LALR parser, unlike machine learning. LALR parser works by following a set of rules created by humans after analyzing patterns on those tagged-tokenized sentences. Finally, in the post-processing stage, those sets of rules or models are assessed, by comparing the model prediction with expert judgment. The resulting evaluation saw the algorithm could parse 86.7% of the evaluation sentences correctly and couldn't parse 10% of the sentences.

Similar to previous research, Mudafiq et al. (Pratama et al., 2017) also use LALR parser to identify sentence structure, the difference is POS tagging was replaced with CFG (Context Grammar). Other than replacing the POS tag with CFG the computing process is largely the same, the algorithm rules are still defined by humans. This research also includes compound sentences in addition to single sentences. The final evaluation shows that LALR parser could successfully parse 70% of the sentence, with the algorithm performing significantly better in a single sentence.

Moving on from LALR, Sulastra et al. used Constraint-Based Formalism to analyze sentence structure in the Indonesian language (Sulastra, 2014). Still the same as the previous two research, in this research algorithm rules are defined by humans. The algorithm successfully parses 61% of the single sentences, and 38% of compound sentences during the evaluation.

Meanwhile, more specifically on the syntactic structure of Indonesian sentences, Wardana uses the Left-Corner algorithm and Nazief and Adriani's stemming process with CSharp programming to categorize sentence functions. Sentences that have been sorted out can be categorized into sentence functions, namely subject (S), predicate (P), object (O), and adverb (K) (Wardana et al., 2019).

Shifting from the Indonesian language, multiple research projects have been conducted to identify sentence structure or grammatical rules in various languages. Such as Arabic (Elarnaoty et al., 2012) and Bangla (Rahman et al., 2020). Nevertheless, research implementing machine learning is almost non-existent, especially in Bahasa Indonesia. Based on this premise, in this research, we created a classification model that is expected to be able to detect the structure of an input sentence.

III. Methods

This research adheres to the SEMMA process, to be more organized and provides a more comprehensive paper. SEMMA process consists of five stages which are sample, explore, modify, model, and assess (Alimyar & Lakshmi G, 2021; Harno et al., 2024; Lima et al., 2024; Novillo Rangone et al., 2021; Quintero et al., 2024; Sundararajan et al., 2020; Tariq et al., 2019; Truong, 2024). The sample stage consisted of gathering the sample, in the explore stage the samples are explored using various data visualization methods, and in the modified stage the sample was manipulated based on the findings in the explore stage, and the model stage saw the sample is fed to the machine learning algorithm, and finally on the assess stage, the machine learning model created on the previous stage was assessed. This flow of process can be repeated if necessary.

A. Sample

One thousand two hundred and fourteen sentences in Bahasa Indonesia had been collected from various resources, both the mass media and intuitive sources based on the reasoning of researchers. Furthermore, the words or phrases in it are labeled according to their position in the sentence structure. The selection of sentence is based on the form of the syntactic structure of Indonesia language sentences, namely subject-predicate (SP); subject-predicate-object (SPO); subject-predicate-object-adverb (SPOK), and subject-predicate-adverb (SPK) (Moeliono et al., 2017). The sentence structure has also been ascertained to be a single sentence in the formal Indonesian language.

B. Explore

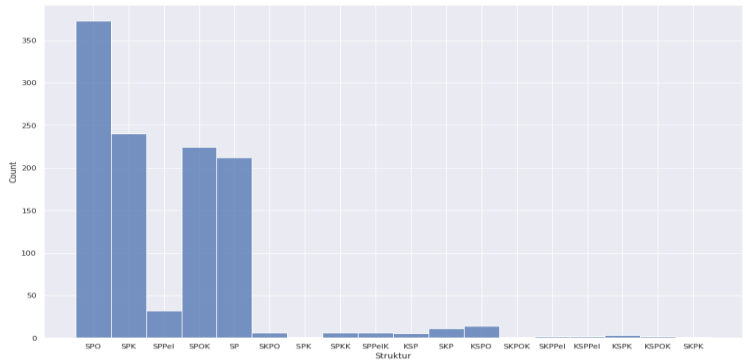

Figure 1. Number of Sentences per Class

Entering the exploration stage, the number of sentences per class is first visualized. Judging from Figure 1, the dataset is imbalanced, thus it is concluded that oversampling is necessary for the next stage. Furthermore, as previously explained, this research only analyzes the structure of SP, SPO, SPOK, and SPK, even though the data contains other sentence structures because the number of other classes is too small.

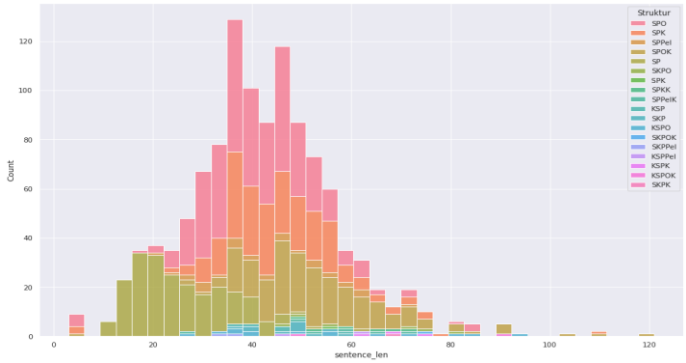

Figure 2. Distribution of Sentence Length

In addition to visualizing the number of each class in the sample, some other visualizations are also conducted. Figure 2 shows the distribution of sentence length which is measured by many characters. From figure 2 it can be concluded that the samples are not evenly distributed, and we can see many sentences consisting of a very small number of characters, in addition to some outliers on the longer sentences. Considering the structure, only SPO and SPK structures are evenly distributed, while SP and SPOK have skewed distribution.

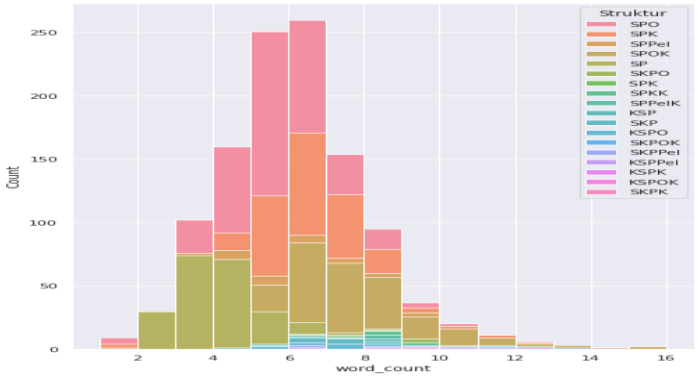

Figure 3. Distribution of Sentence Length

Figure 3 shows the distribution of word counts of the samples. Again, the samples are not evenly distributed, but contrary to Figure 2, only the outliers on the longer sentences caused the skewness. Structure-wise, the SP and SPOK still have a noticeable difference in distribution. This is because SP sentences are mostly very short.

C. Modify

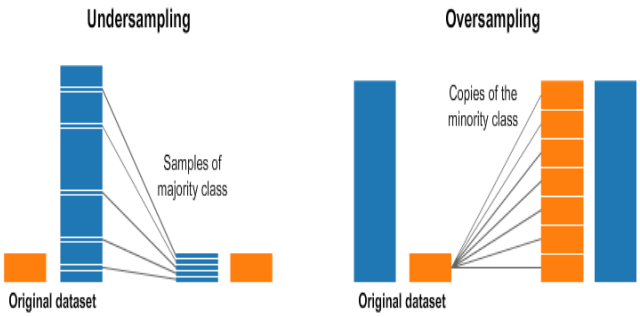

Figure 4. Comparison of Undersampling and Oversampling

In the previous stage, we found that the sample was imbalanced. An imbalanced sample could impact the accuracy of the model negatively. There are two resampling methods to handle an imbalanced sample, under-sampling and oversampling, which Figure 4 shows the comparison. Under-sampling took the majority class, which in this research are SPO, SPK, and SPOK, and then randomly removed sentences on those structures until the number of sentences was equal to the number of sentences on SP. Oversampling took the minority class, which is SPK, SPOK, and SP, and then multiplied some sentences until the number of sentences was equal to the number of sentences on SPO (Table 1).

TABLE 1
NUMBER OF SENTENCES PER STRUCTURE

Structure	Number of Sentences		
	Sample	***Undersampled***	***Oversampled***
SPO	373	212	373
SPK	240	212	373
SPOK	224	212	373
SP	212	212	373

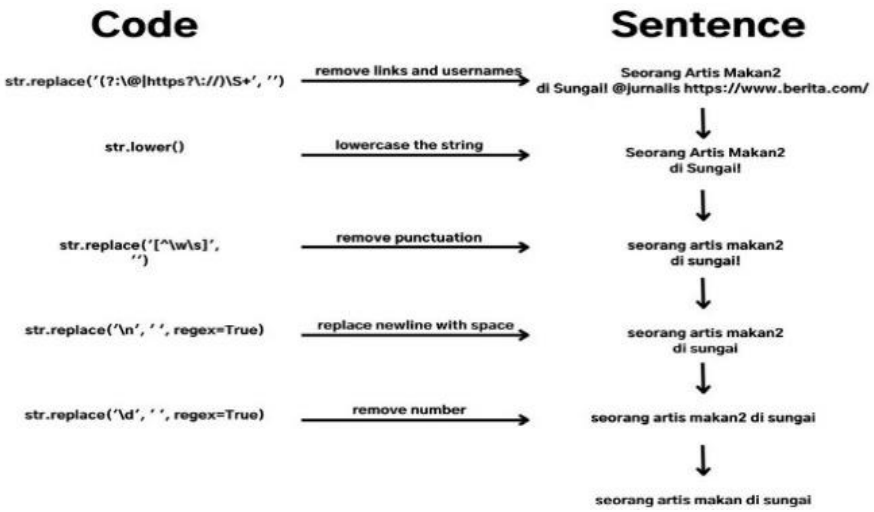

Figure 5. Data Cleaning Process of This Research

It is important to note that the samples are collected from various sources, which might contain unwanted characters. Thus, the sentences on the sample undergo a series of data-cleaning syntaxes. Figure 5 shows the syntaxes used and the effects of each syntax.

Contrary to popular belief, computers cannot directly understand human language. Human language must undergo a series of processes to be understood by computers; there are a myriad of processes but all of them come down to turning the text into numbers. This process is known as vectorization. The experimentation in this research will focus on implementing 4 different vectorization methods and compare the accuracy of the resulting model in predicting the sentence structure.

Sample

Sentence	Structure
Saya minum	SP
Saya makan nasi	SPK
Saya makan nasi di Jonggol	SPOK

Computer would process as

ayam	di	jonggol	makan	minum	nasi	saya	Class
0	0	0	0	1	0	1	0
0	0	0	1	0	1	1	1
1	1	1	1	0	0	1	2

Figure 6. Bag of Words Vectorization

One of the used vectorization methods is Bag of Words, the simplest of all. Bag of Words or BoW essentially turns every word that exists on the sample into a feature. Figure 6 illustrates the results of BoW. However, there is one crucial weakness which is the fact that BoW doesn't consider the position of a word in a sentence or the context of the word.

Bag-of-Words (BoW) is a simple yet powerful technique for representing text data in the form of numerical feature vectors. The basic idea behind BoW is to treat a document as a collection of its words, disregarding grammar and word order, and represent it as a fixed-length vector where each dimension corresponds to a specific word and the value in that dimension represents the number of times that word appears in the document. BoW is commonly used in natural language processing tasks such as text classification, sentiment analysis, and topic modeling.

One famous research that used the BoW approach is "A Study on Text Classification Using SVM and Naive Bayes" by Joachims, T. (Joachims, 1999), where the authors applied SVM and Naive Bayes classification algorithms to the text classification problem using the BoW representation. Another research is "A Comparative Study on Feature Selection in Text Categorization" by Yang, Y. (Yang et al., 2012) where the authors used BoW representation as a feature for text categorization and compared it with other feature selection methods. BoW is still widely used as feature representation in many text mining tasks, and it is effective in many real-world applications.

Another common method of vectorization is Term Frequency-Inverse Document Frequency, otherwise known as TF-IDF. This method partially counters BoW weakness by replacing the numbers, instead of just 0 and 1, TF-IDF adheres to the following formulas.

$$idf_i = \log(\frac{n}{df_i}) \quad (1)$$

$$w_{i,j} = tf_{i,j} \times idf_i \quad (2)$$

To put this formula in layman's terms. TF-IDF is almost like BoW with the exception that $w_{i,j}$ would replace the cells that have the value one, $tf_{i,j}$ is the same as BoW. Hence TF-IDF is multiplying idf_i value of each word with a BoW (Salton & Buckley, 1988), as shown in Figure 7 by using a sample previously used in Figure 6.

Implement formula (1) for every word on sample

Word	Idf_i
ayam	Log(1/3) = 0.477
di	Log(1/3) = 0.477
jonggol	Log(1/3) = 0.477
makan	Log(2/3) = -0.17
minum	Log(1/3) = 0.477
nasi	Log(1/3) = 0.477
saya	Log(3/3) = 0

$tf_{i,j}$ (Similar to BoW)

ayam	di	jonggol	makan	minum	nasi	saya	Class
0	0	0	0	1	0	1	0
0	0	0	1	0	1	1	1
1	1	1	1	0	0	1	2

Implement formula (2), computer would process as:

ayam	di	jonggol	makan	minum	nasi	saya	Class
0	0	0	0	0.477	0	0	0
0	0	0	-0.17	0	0.477	0	1
0.477	0.477	0.477	-0.17	0	0	0	2

Figure 7. TF-IDF Vectorization

There are some issues with the BoW and TF-IDF approach. Firstly, the measure of vectors depends on the estimate of our lexicon, which can be colossal. Usually, this is a squander of space and increments algorithms complexity exponentially which would result in *The Curse of Dimensionality*.

Furthermore, these embeddings will be closely coupled to their applications, making transfer-learning to a model employing a diverse lexicon of the same estimate, including, or expelling words from vocabulary would be nearly outlandish because it would require to re-train the complete model once more.

Finally, the complete reason for embedding is to capture the relevant meaning of the words, which this representation comes up short to do. There's no correlation between words that have comparable meaning or utilization. To tackle the problems, we mentioned CBoW and Skip-Gram were implemented.

The first Skip-gram model is defined as a set of assembled word prediction assignments. Each assignment comprises of prediction of a word v given a word w utilizing correspondingly their yield and input representations.

$$p(v \mid w, \theta) = \frac{\exp\left(in_w^{\mathrm{T}} \text{ out }_v\right)}{\sum_{v'=1}^{V} \exp\left(in_w^{\mathrm{T}} \text{ out } t_{v'}\right)}, \tag{3}$$

where global parameter $\theta = \{in_v, \text{ out }_v\}_{v=1}^{V}$ stands for both input and output representations for all words of the dictionary indexed with $1, \ldots, V$. Both input and output representations are real vectors of the dimensionality D. These individual predictions are grouped in a way to simultaneously predict context words y of some input word x.

$$p(\mathbf{y} \mid x, \theta) = \prod_j p\left(y_j \mid x, \theta\right). \tag{4}$$

Input text $\mathbf{o}$ consisting of N words $o_1, o_2, \ldots, o_N$ is then interpreted as a sequence of input words $X = \{x_i\}_{i=1}^{N}$ and their contexts $Y = \{\mathbf{y}_i\}_{i=1}^{N}$. Here i-the training object $(x_i, \mathbf{y}_i)$ consists of the word $x_i = o_i$ and its context $\mathbf{y}_i = \{o_t\}_{t \in c(i)}$ where $c(i)$ is a set of indices such that $|t - i| \leq C/2$ and $t \neq i$ for all $t \in c(i)$.

Finally, the Skip-gram objective function is the likelihood of contexts given the corresponding input words:

$$p(Y \mid X, \theta) = \prod_{i=1}^{N} p(\mathbf{y}_i \mid x_i, \theta) = \prod_{i=1}^{N} \prod_{j=1}^{C} p\left(y_{ij} \mid x_i, \theta\right) \tag{5}$$

(Mikolov et al., 2013)

A continuous bag of words model (CBOW) is an exceedingly effective shallow neural network algorithm. It is created to produce vector representations of a language lexicon so that the information of the words is encoded within the vector space structure. The continuous bag of words model is comparable to the feedforward neural network language model (FFNNLM). FFNNLM was first proposed by Bengio et al., which applies the feedforward neural arrange (FFNN) into the language model and learns a dispersed representation for words to unravel the issue of high dimensionality. Be that as it may, the non-linear hidden layer of the model is erased, and the projection layer of the model is applied to all words. That's to say, the continuous-bag-of-words model may be a basic neural network with three layers: input, projection, and output (Mikolov et al., 2013).

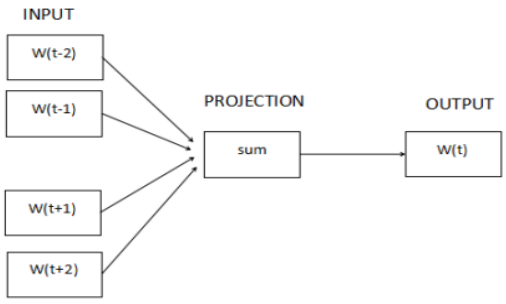

Figure 8. Continuous Bag of Words

D. Model

After undergoing vectorization, the dataset would move on to the next phase, which is the modeling phase. In this phase, the vectorized dataset will be modeled with the Random Forest algorithm. The random forest classifier comprises a combination of tree classifiers where each classifier is created employing a random vector tested freely from the input vector and each tree casts a unit vote for the foremost well-known class to classify an input vector. Random forests classifier is one of the foremost fruitful ensemble learning procedures that have been demonstrated to be exceptionally prevalent and effective in pattern recognition and machine learning for high-dimensional classification and skewed problems. A disadvantage related to tree classifiers is their high variance. In practice, it isn't exceptional for a little alteration within the training data set to result in an exceptionally diverse tree (Breiman, 2001).

E. Assess

Models created during the previous stage are then assessed in this stage. The assessment uses a standard practice for assessing classification models which include the use of Confusion Matrix. A confusion matrix is a table that is used to

define the performance of a classification algorithm. It is a summary of the predictions made by a classifier in comparison to the true values. Confusion matrix is a powerful tool for measuring the quality of classification models.

The basic mathematical formula for a confusion matrix is a table with rows representing the true values and columns representing the predicted values. For example, in a binary classification problem, the matrix will have two rows and two columns. The cells in the matrix represent the number of observations that have been predicted to belong to a certain class while they belong to another class. The diagonal elements represent the number of observations that have been correctly classified, while the off-diagonal elements represent the number of observations that have been misclassified.

Mosteller and Wallace first introduced the confusion matrix in a 1964 paper "Inference in an Authorship Problem" (Mosteller & Wallace, 1963). Confusion matrix is widely used in many machine learning tasks for evaluating the performance of a classification model, it is also used to evaluate the performance of other models that make categorical predictions, such as decision trees, neural networks, and support vector machines.

		Predicted Value			
		SP	SPO	SPK	SPOK
Real Value	SP	P-SP-SP	P-SPO-SP	P-SPK-SP	P-SPOK-SP
	SPO	P-SP-SPO	P-SPO-SPO	P-SPK-SPO	P-SPOK-SPO
	SPK	P-SP-SPK	P-SPO-SPK	P-SPK-SPK	P-SPOK-SPK
	SPOK	P-SP-SPOK	P-SPO-SPOK	P-SPK-SPOK	P-SPOK-SPOK

Figure 9. The Confusional Matrix of This Research

IV. FINDINGS

In this study, eight different experiments have been conducted to explore the effects of different vectorization techniques and sampling methods for classifying Indonesian sentences based on their structure, focusing on four basic sentence types: SP, SPO, SPK, and SPOK. The dataset used for these experiments was approximately 1214 labeled sentences, the entire experiment used the Random Forest algorithm as the classifier. There are four vectorization methods tested which include Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), Continuous Bag of Words (CBoW), and Skip-gram, in addition to oversampling and under-sampling is also used to address class imbalance. Figure 10 below displays the summary of the experiment results.

Experiment Number	Sampling	Vectorization	Accuracy
1	Oversampled	BoW	69%
2	Undersampled	BoW	65%
3	Oversampled	TF-IDF	70%
4	Undersampled	TF-IDF	73%
5	Oversampled	CBoW	76%
6	Undersampled	CBoW	68%
7	Oversampled	Skipgram	68%
8	Undersampled	Skipgram	65%

Figure 10. Experiment Results

The CBoW model combined oversampling achieved the highest accuracy (76%), indicating it has the best capability in classifying sentence structures. This finding is likely due to how CBoW works itself, in which the model predicts a target word based on its surrounding context that contributes to creating a representation of sentence structures. The efficiency of CBoW in encoding sentence context likely contributed to its better performance in distinguishing between the complex sentence structures represented in the dataset, as we can see from its confusion matrix shown in Figure 11.

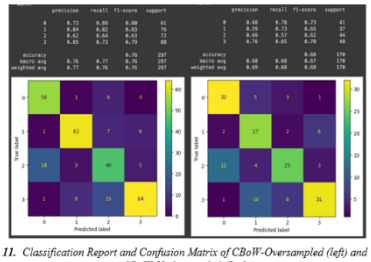

Fig. 11. Classification Report and Confusion Matrix of CBoW-Oversampled (left) and CBoW-Undersampled (Right)

In contrast, the Skip-gram model, despite functioning by creating a detailed word relationship mapping, showed lower accuracies, which are 68% with oversampling and 65% with under-sampling. This can be attributed to Skip-gram's focus on predicting surrounding words for a given target word, which, while useful for capturing word meanings, may not be as effective for the task of sentence structure classification. As we can see from its confusional matrix in Figure 12, Skip-gram fails in distinguishing between SPK and SPOK, this is likely due to its detail-oriented structures. This is better at capturing relationships involving rare words or phrases, but this detail-oriented approach might not be as necessary for understanding general sentence structures, especially when the focus is on common syntactic patterns.

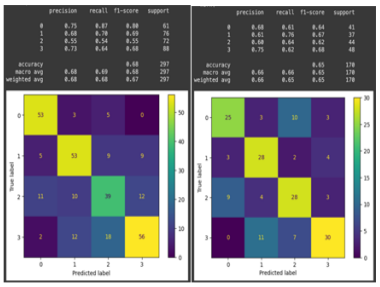

Fig. 12. Classification Report and Confusion Matrix of Skipgram-Oversampled (left) and Skipgram-Undersampled (Right)

On the other hand, BoW and TF-IDF might be too simple for this task. In the case of BoW, the oversampling method might have helped to balance the classes, leading to better model performance compared to under-sampling. However, BoW does not consider the order of words or context, which might limit its effectiveness in capturing sentence structures. Compared to oversampling, the decrease in accuracy in BoW_under-sampling could be due to the loss of information caused by under-sampling itself. This method might have removed valuable examples, leading to a less effective training process for the model.

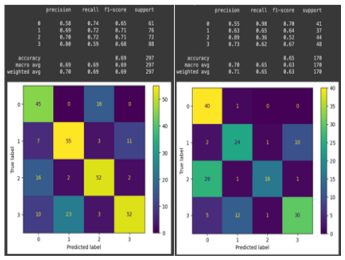

Fig. 13.Classification Report and Confusion Matrix of BoW-Oversampled (left) and BoW-Undersampled (Right)

However, in the case of TF-IDF, things are reversed. This method weighs the words based on their frequency across the datasets, which can provide better insights into each word concerning sentence structures. The impact of sampling methods on model performance was also notable. TF-IDF with under-sampling achieved higher accuracy (73%) than with oversampling (70%), suggesting that the model could generalize better with a more balanced, albeit smaller, dataset. This result underscores the potential trade-off between dataset size and balance in machine learning, where reducing the number of samples to achieve class balance can sometimes lead to more robust models.

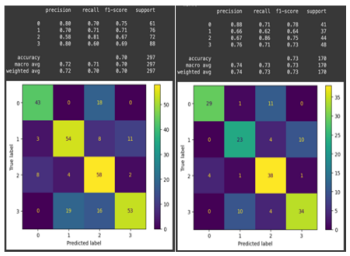

Fig. 14. Classification Report and Confusion Matrix of TFIDF-Oversampled (left) and TFIDF-Undersampled (Right)

V. Discussion

This study explored the impact of various vectorization techniques and sampling methods on classifying Indonesian sentences based on their structure using the Random Forest algorithm. Four basic sentence types were classified: Subject-Predicate (SP), Subject-Predicate-Object (SPO), Subject-Predicate-Complement (SPK), and Subject-Predicate-Object-Complement (SPOK). Through eight different experiments, we aimed to determine the most effective vectorization and sampling combinations for achieving high classification accuracy.

A. Vectorization Techniques

The experiments tested four vectorization methods: Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), Continuous Bag of Words (CBoW), and Skip-gram. Each method offers a unique approach to representing text data, impacting how well sentence structures can be distinguished by the classifier.

The results show that *CBoW combined with oversampling achieved the highest accuracy (76%)*, making it the most effective method in this study. CBoW works by predicting a target word based on its surrounding context, enabling it to capture sentence structure better than other methods. This efficiency in encoding the contextual relationship between words allows CBoW to more accurately differentiate between the four sentence types. The confusion matrix in Figure 11 supports this, as CBoW-oversampling shows fewer misclassifications, especially between closely related sentence structures such as SPO and SPOK.

In contrast, the *Skip-gram model yielded lower accuracy* (68% with oversampling and 65% with undersampling), which can be explained by its mechanism of predicting surrounding words from a given target word. Although Skip-gram excels in capturing detailed word relationships, it may not be as effective in sentence structure classification, where understanding the general syntactic patterns is more important than detailed word-level relations. As indicated in the confusion matrix (Figure 12), Skip-gram particularly struggled to distinguish between SPK and SPOK, likely because its focus on rare word relationships does not contribute much to identifying common sentence structures.

The *BoW and TF-IDF models* both performed reasonably well but with notable differences in their performance. BoW with oversampling achieved 69% accuracy, slightly higher than undersampling (65%). This is likely because oversampling helped balance the classes, allowing the model to train on a more evenly distributed dataset. However, BoW's limitation lies in its inability to capture the word order or context, which is essential for understanding sentence structure. The underperformance with undersampling can be attributed to the loss of critical information, as valuable training examples may have been removed, reducing the model's effectiveness.

In the case of *TF-IDF*, under-sampling produced a better result (73% accuracy) than oversampling (70%). This suggests that under-sampling helped the model generalize better, even with a smaller dataset. TF-IDF considers word frequency across the dataset, which allows it to weigh words in a way that better reflects sentence structure. This finding highlights the importance of balancing dataset size and quality in machine learning models, where fewer but more balanced samples can sometimes lead to more robust performance, as evidenced by the confusion matrix in Figure 14.

B. Impact of Sampling Methods

Sampling methods, particularly oversampling and undersampling, were employed to address the class imbalance in the dataset. The impact of these methods was evident across all experiments. Generally, oversampling improved the model performance for vectorization methods like CBoW and BoW, as it mitigated the issue of underrepresented classes. However, oversampling also risked introducing noise by duplicating samples, which may explain why TF-IDF performed better with under-sampling.

Undersampling, on the other hand, produced mixed results. For TF-IDF, it resulted in better model accuracy, likely because it removed noisy or redundant data, allowing the classifier to focus on more representative examples of each sentence type. However, for methods like BoW and Skip-gram, undersampling led to a decrease in accuracy, suggesting that these methods require a larger dataset to capture enough information about sentence structures.

C. Implications for Indonesian Sentence Classification

The findings of this study have significant implications for classifying Indonesian sentence structures. First, the superiority of CBoW in combination with oversampling demonstrates that capturing the contextual relationships between words is critical for understanding sentence structures. This suggests that future studies should explore further enhancements to context-based models, such as more sophisticated embeddings or deep learning techniques like BERT, which could potentially improve classification accuracy.

Moreover, the trade-offs between oversampling and undersampling highlight the importance of dataset balance in machine learning. While oversampling can help address class imbalances, it may not always lead to improved performance, particularly for vectorization methods like TF-IDF, which benefit from undersampling. This finding suggests that achieving optimal performance may require fine-tuning the balance between dataset size and quality, depending on the specific vectorization method employed.

Finally, the lower performance of BoW and Skip-gram indicates that these methods may not be well-suited for sentence structure classification tasks that require an understanding of the syntactic arrangement of words. Future research should consider exploring more advanced models that can capture word order and context more effectively, such as recurrent neural networks (RNNs) or transformers.

VI. Conclusion

The study's results highlight the importance of choosing appropriate vectorization and sampling techniques for specific linguistic tasks. CBoW's context-capturing efficiency, especially when combined with oversampling, proved to be most effective for classifying Indonesian sentence structures, offering a promising direction for further research in computational linguistics and natural language processing. The linguistic characteristics of the Indonesian language, with

its structured sentence construction (SP, SPO, SPK, SPOK), played a significant role in the experiment outcomes. The ability of CBoW to capture sentence-wide context likely provided it with an advantage in identifying the underlying structure of sentences, which is less pronounced in models like Skip-gram that emphasize word-to-word relationships.

While the findings offer valuable insights into the suitability of different vectorization and sampling methods for sentence structure classification in Indonesian, the study has limitations, including the relatively small size of the dataset. Future research could explore larger datasets, different machine learning algorithms, and deep learning approaches to further understand the dynamics of sentence structure classification in Indonesian and other languages with similar syntactic features.

REFERENCES

[1] Abeywickrama, P., & Brown, H. (2010). *Language Assessment: Principles and Classroom Practices.* NY: Pearson Longman.

[2] Ahmad, S. N., & Laroche, M. (2023). Extracting Marketing Information from Product Reviews: A Comparative Study of Latent Semantic Analysis and Probabilistic Latent Semantic Analysis. *Journal of Marketing Analytics*, *11*(4), 662-676. https://doi.org/10.1057/s41270-023-00218-6

[3] Alimyar, Z., & Lakshmi, G, S. (2021). A Study on Language Teachers' Preparedness to Use Technology during COVID-19. *Cogent Arts & Humanities*, *8*(1), 1999064. https://doi.org/10.1080/23311983.2021.1999064

[4] Alimyar, Z., & Lakshmi G, S. (2021). A Study on Language Teachers' Preparedness to Use Technology during COVID-19. *Cogent Arts and Humanities*, *8*(1), 1999064. https://doi.org/10.1080/23311983.2021.1999064

[5] Arooj, S., Altaf, S., Ahmad, S., Mahmoud, H., & Mohamed, A. S. N. (2024). Enhancing sign language recognition using CNN and SIFT: A case study on Pakistan sign language. *Journal of King Saud University-Computer and Information Sciences*, *36*(2), 101934. https://doi.org/10.1016/j.jksuci.2024.101934

[6] Breiman, L. (2001). Random Forests. *Machine learning*, *45*, 5-32. https://doi.org/10.1023/A:1010933404324

[7] Campbell, M. I., Rai, R., & Kurtoglu, T. (2012). A Stochastic Tree-Search Algorithm for Generative Grammars. *Journal of Computing and Information Science in Engineering*, *12*(3), 031006. https://doi.org/10.1115/1.4007153

[8] De Smedt, F., Van Keer, H., & Merchie, E. (2016). Student, Teacher and Class-Level Correlates of Flemish Late Elementary School Children's Writing Performance. *Reading and Writing*, *29*, 833-868. https://doi.org/10.1007/s11145-015-9590-z

[9] Elarnaoty, M., AbdelRahman, S., & Fahmy, A. (2012). A Machine Learning Approach for Opinion Holder Extraction in Arabic Language. *arXiv preprint arXiv:1206.1011*. https://doi.org/10.5121/ijaia.2012.3205

[10] Fatonah, K., & Wiradharma, G. (2018). *Pemetaan Genre Teks Bahasa Indonesia pada Kurikulum 2013 (Revisi) Jenjang SMA* [Mapping of Indonesian Language Text Genres in the 2013 Curriculum (Revised) for High School Level]. https://repositori.kemdikbud.go.id/10046/1/dokumen_makalah_1540362989.pdf (taken d/d March 19, 2025)

[11] Friedman, J. (1969). A Computer System for Transformational Grammar. *Communications of the ACM*, *12*(6), 341-348. https://doi.org/10.1145/363011.363154

[12] Friedman, J. (1971). *A Computer Model of Transformational Grammar*. https://doi.org/10.1145/363011.363154

[13] Gunawan, D., Siregar, H. P., & Sitompul, O. S. (2019). Identifying Sentence Structure in Bahasa Indonesia by Using POS Tag and LALR Parser. *2019 5th International Conference on Computing Engineering and Design (ICCED)*.

[14] Hariyanto, P., Zalmansyah, A., Endardi, J., Sukesti, R., Sumadi, S., Abidin, Z., . . . Ratnawati, R. (2023). Language maintenance and identity: A case of Bangka Malay. *International Journal of Society, Culture, and Language*, *11*(2 (Themed Issue on Language, Discourse, and Society)), 60-74. https://doi.org/10.22034/ijscl.2023.2002013.3030

[15] Harno, S., Chan, H. K., & Guo, M. (2024). Enhancing Value Creation of Operational Management for Small to Medium Manufacturer: A Conceptual Data-Driven Analytical System. *Computers and Industrial Engineering*, *190*, 110082. https://doi.org/10.1016/j.cie.2024.110082

[16] Hoyos Pipicano, Y. A. (2024). Exploring Standardized Tests Washback from the Decolonial Option: Implications for Rural Teachers and Students. *Cogent Arts and Humanities*, *11*(1), 2300200. https://doi.org/10.1080/23311983.2023.2300200

[17] Jagaiah, T., Olinghouse, N. G., & Kearns, D. M. (2020). Syntactic Complexity Measures: Variation by Genre, Grade-Level, Students' Writing Abilities, and Writing Quality. *Reading and Writing*, *33*, 2577-2638. https://link.springer.com/article/10.1007/s11145-020-10057-x

[18] Jiang, L., Yu, S., & Lee, I. (2022). Developing A Genre-Based Model for Assessing Digital Multimodal Composing in Second Language Writing: Integrating Theory with Practice. *Journal of Second Language Writing*, *57*, 100869. https://doi.org/10.1016/j.jslw.2022.100869

[19] Joachims, T. (1999). Transductive inference for text classification using support vector machines. *Proceedings of the 20th International Conference on Machine Learning*.

[20] Khair, U., & Misnawati, M. (2022). Indonesian Language Teaching in Elementary School: Cooperative Learning Model Explicit Type Instructions Chronological Technique of Events on Narrative Writing Skills from Interview Texts. *Linguistics and Culture Review*, 172-184. https://doi.org/10.21744/lingcure.v6nS2.1974

[21] Kim, S., Park, H., & Lee, J. (2020). Word2vec-Based Latent Semantic Analysis (W2V-LSA) for Topic Modeling: A Study on Blockchain Technology Trend Analysis. *Expert Systems With Applications*, *152*, 113401. https://doi.org/10.1016/j.eswa.2020.113401

[22] Kim, Y. S. G., & Zagata, E. (2024). Enhancing Reading and Writing Skills through Systematically Integrated Instruction. *The Reading Teacher*, *77*(6), 787-799. https://doi.org/10.1002/trtr.2307

[23] Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2011). *Automated Grammatical Error Detection for Language Learners.* https://doi.org/10.1162/COLI_r_00062

[24] Lee, M. C., Chang, J. W., & Hsieh, T. C. (2014). A Grammar-Based Semantic Similarity Algorithm for Natural Language Sentences. *The Scientific World Journal*, *2014*(1), 437162. https://doi.org/10.1155/2014/437162

[25] Lima, J. F., Acosta-Urigüen, M. I., & Orellana, M. (2024). Machine learning and knowledge engineering for cognitive memory assessment of age groups by anomalies in a serious game. *Intelligent Systems With Applications*, *21*, 200301. https://doi.org/10.1016/j.iswa.2023.200301
[26] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space.* arXiv preprint arXiv:1301.3781. https://doi.org/10.48550/arXiv.1301.3781
[27] Moeliono, A. M., Lapoliwa, H., Alwi, H., Tjatur, S. S., Sasangka, W., & Sugiyono, S. (2017). *Tata bahasa baku bahasa Indonesia. Edisi keempat* [Standard grammar of Indonesian language. Fourth edition] https://repositori.kemdikbud.go.id/16351/ (taken d/d March 19, 2025)
[28] Moeljadi, D., Sugianto, R., Hendrick, J. S., & Hartono, K. (2016). *Kamus Besar Bahasa Indonesia (KBBI)* [Indonesian Dictionary (KBBI)]. *Badan Pengembangan Bahasa dan Kebukuan, Kementerian Pendidikan dan Kebudayaan*. https://davidmoeljadi.github.io/slides/kbbi2.pdf (taken d/d March 19, 2025)
[29] Mosteller, F., & Wallace, D. L. (1963). Inference in an Authorship Problem: A Comparative Study of Discrimination Methods Applied to the Authorship of the Disputed Federalist Papers. *Journal of the American Statistical Association*, *58*(302), 275-309. https://doi.org/10.1080/01621459.1963.10500849
[30] Nardiati, S., Isnaeni, M., Widodo, S. T., Hardaniwati, M., Susilawati, D., Winarti, S., ... Zalmansyah, A. (2023). Cultural and philosophical meaning of Javanese traditional houses: A case study in Yogyakarta and Surakarta, Indonesia. *Eurasian Journal of Applied Linguistics*, *9*(2), 1-10. https://ejal.info/menuscript/index.php/ejal/article/view/516 (taken d/d March 19, 2025)
[31] Nordin, N. R. M., Omar, W., & Ridzuan, I. N. I. M. (2022). Challenges and Solutions of Online Language Teaching and Assessment during COVID-19. *World Journal of English Language*. https://doi.org/10.5430/wjel.v12n8p410
[32] Novillo Rangone, G., Pizarro, C., & Montejano, G. (2021). Automation of an Educational Data Mining Model Applying Interpretable Machine Learning and Auto Machine Learning. *International Conference on Communication and Applied Technologies*, 22-30. https://doi.org/10.1007/978-981-16-5792-4_3
[33] Nursugiharti, T., Zalmansyah, A., & Rasyid, F. M. (2024). *Religious Values of the Traditional Ceremony in Building a Bengkulu Malay Traditional House*. In: ISVS e-journal.
[34] Philippakos, Z. A. T., MacArthur, C. A., & Rocconi, L. M. (2023). Effects of Genre-Based Writing Professional Development on K to 2 Teachers' Confidence and Students' Writing Quality. *Teaching and Teacher Education*, *135*, 104316. https://doi.org/10.1016/j.tate.2023.104316
[35] Pratama, M. R., Kusumadewi, S., & Hidayat, T. (2017). Penerapan Algoritma Lalr Parser dan Context-Free Grammar untuk Struktur Kalimat Bahasa Indonesia [Application of the Lalr Parser Algorithm and Context-Free Grammar for Indonesian Sentence Structure]. *Jurnal Teknologi Elektro*, *8*(1), 1-8. https://doi.org/10.22441/jte.v8i1.1364
[36] Quintero, J. B., Villanueva-Valdes, D., & Manrique-Losada, B. (2024). Artificial Neural Networks in the Development of Business Analytics Projects. *International Journal of Information and Decision Sciences*, *16*(1), 46-72. https://doi.org/10.1504/IJIDS.2024.136283
[37] Rabiah, S. (2018). *Language as A Tool for Communication and Cultural Reality Discloser.* http://dx.doi.org/10.31227/osf.io/nw94m
[38] Rahman, M., Haque, S., & Saurav, Z. R. (2020). Identifying and Categorizing Opinions Expressed in Bangla Sentences Using Deep Learning Technique. *International Journal of Computer Applications*, *176*(17), 13-17. https://doi.org/10.5120/ijca2020920119
[39] Ramliyana, R., Pratiwi, N. K., & Megiati, Y. E. (2022). Analysis of Indonesian Language Error in Writing Reports of Students' Learning Results of The Amanah Fitrah Rabbani Foundation Using The Sipebi Application. *Hortatori: Jurnal Pendidikan Bahasa dan Sastra Indonesia*, *6*(1), 6-16. https://doi.org/10.30998/jh.v6i1.998
[40] Ratna, A. A. P., Purnamasari, P. D., & Adhi, B. A. (2015). *SIMPLE-O, the Essay Grading System for Indonesian Language Using LSA Method with Multi-Level Keywords.* The Asian Conference on Society, Education & Technology.
[41] Renza, M. A., Affandi, L. H., & Setiawan, H. (2022). Pengembangan Media Gambar Berseri pada Materi Keterampilan Menulis Teks Narasi Siswa Kelas IV. *Jurnal Ilmiah Profesi Pendidikan*, *7*(2), 445-451. https://doi.org/10.29303/jipp.v7i2.562
[42] Rodríguez-Gonzalo, C., & Abad-Beltrán, V. (2023). Teaching Writing through Discourse Genres. In *Development of writing skills in children in diverse cultural contexts: contributions to teaching and learning* (pp. 301-323). Springer. https://doi.org/10.1007/978-3-031-29286-6_14
[43] Salton, G., & Buckley, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval. *Information processing & management*, *24*(5), 513-523. https://doi.org/10.1016/0306-4573(88)90021-0
[44] Setiawan, D., Hartati, T., & Sopandi, W. (2019). Kemampuan Menulis Teks Eksplanasi Siswa Kelas 5 Sekolah Dasar melalui Model Read, Answer, Disscuss, Explain, And Create: Radec. *Pendas: Jurnal Ilmiah Pendidikan Dasar*, *4*(1), 1-16. https://doi.org/10.23969/jp.v4i1.1575
[45] Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated Essay Scoring: Writing Assessment and Instruction. *International Encyclopedia of Education*, *4*(1), 20-26. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=eed567622453f29b7b2d72955d07d8aad5e86daf (take d/d March 19, 2025)
[46] Sinaga, T., Kadaryanto, B., & Aulia, N. (2023). Indonesian High School Students' Critical Thinking and Literary Text Comprehension. *ELE Reviews: English Language Education Reviews*, *3*(2), 155-171. https://doi.org/10.22515/elereviews.v3i2.7621
[47] Sulastra, J. (2014). Perancangan penganalisis struktur kalimat bahasa indonesia dengan menggunakan constraint-based formalism. *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi*, *5*(2), 1–11.
[48] Sun, T., Wang, C., & Wang, Y. (2022). The Effectiveness of Self-Regulated Strategy Development on Improving English Writing: Evidence from the Last Decade. *Reading and Writing*, *35*(10), 2497-2522. https://doi.org/https://doi.org/10.1007/s11145-022-10297-z
[49] Sundararajan, A., Hernandez, A. S., & Sarwat, A. I. (2020). Adapting Big Data Standards, Maturity Models to Smart Grid Distributed Generation: Critical Review. *IET Smart Grid*, *3*(4), 508-519. https://doi.org/https://doi.org/10.1049/iet-stg.2019.0298

[50] Tariq, H. I., Sohail, A., Aslam, U., & Batcha, N. K. (2019). Loan Default Prediction Podel using Sample, Explore, Modify, Model, and Assess (SEMMA). *Journal of Computational and Theoretical Nanoscience*, *16*(8), 3489-3503. http://dx.doi.org/10.1166/jctn.2019.8313
[51] Tektigul, Z., Bayadilova-Altybayev, A., Sadykova, S., Iskindirova, S., Kushkimbayeva, A., & Zhumagul, D. (2023). Language is a symbol system that carries culture. *International Journal of Society, Culture & Language*, *11*(1), 203-214. https://doi.org/10.22034/ijscl.2022.562756.2781
[52] Truong, D. (2024). *Data Science and Machine Learning for Mon-Programmers: using SAS Enterprise Miner.* https://doi.org/10.1080/00401706.2024.2374190
[53] Utami, S. P. T. (2022). *Teknologi dalam Penyuntingan Naskah Bahasa Indonesia: Studi Komparasi Pemanfaatan Aplikasi SIPEBI, Ejaan. id, lektur. id, typoonline. com, dan typograp. com* [Technology in Editing Indonesian Manuscripts: A Comparative Study of the Utilization of SIPEBI Applications, Ejaan.id, lektur.id, typoonline.com, and typograp.com]. *Itell*. https://itell.or.id/conference/index.php/itell/itell2022/paper/view/169 (taken d/d March 19, 2025)
[54] Valdez, D., Pickett, A. C., & Goodson, P. (2018). Topic Modeling: Latent Semantic Analysis for the Social Sciences. *Social Science Quarterly*, *99*(5), 1665-1679. https://doi.org/10.1111/ssqu.12528
[55] Venugopal, M., Sharma, V. K., & Sharma, K. (2023). Web Information Mining and Semantic Analysis in Heterogeneous Unstructured Text Data Using Enhanced Latent Dirichlet Allocation. *Concurrency and Computation: Practice and Experience*, *35*(1), e7410. https://doi.org/10.1002/cpe.7410
[56] Wardana, H. K., Swanita, I., & Yohanes, B. W. (2019). Sistem Pemeriksa Pola Kalimat Bahasa Indonesia berbasis Algoritme Left-Corner Parsing dengan Stemming. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, *8*(3), 211-217. https://doi.org/10.22146/jnteti.v8i3.515
[57] Winahyu, S. K. (2024). *Pengembangan Instrumen Penilaian Keterampilan Menulis Artikel Opini Bahasa Indonesia Berbasis Komputer* [Development of Computer-Based Indonesian Language Opinion Article Writing Skills Assessment Instrument] [UNIVERSITAS NEGERI JAKARTA]. https://lib.unj.ac.id/tugasakhir/index.php?p=show_detail&id=85003 (taken d/d March 19, 2025)
[58] Wortmann, T., & Stouffs, R. (2018). Algorithmic Complexity of Shape Grammar Implementation. *AI EDAM*, *32*(2), 138-146. https://doi.org/https://doi.org/10.1017/S0890060417000440
[59] Yang, Y., Hua, J. X., Xin, H. D., & Li, X. (2012). *Comparative Study on Feature Selection in Uighur Text Categorization.* 19-26. https://doi.org/10.4156/AISS.vol4.issue3.3
[60] Zalmansyah, A. (2017). Meningkatkan Perbendaharaan Kata (*Vocabulary*) Siswa dengan Menggunakan Komik Strip sebagai Media Pembelajaran Bahasa Inggris. *Kandai*, *9*(2), 262-275. doi: https://doi.org/10.26499/jk.v9i2.292
[61] Zalmansyah, A. (2017). *Meningkatkan Perbendaharaan Kata (Vocabulary) Siswa dengan Menggunakan Komik Strip sebagai Media Pembelajaran Bahasa Inggris* [Improving Students' Vocabulary by Using Comic Strips as English Learning Media] *Kandai*, *9*(2), 262-275. https://doi.org/10.26499/jk.v9i2.292
[62] Zalmansyah, A. (2018). *Teknik Cooperative Integrated Reading and Composition (CIRC) untuk Meningkatkan Kemampuan Menulis* [Cooperative Integrated Reading and Composition (CIRC) Technique to Improve Writing Skills] *Ranah: Jurnal Kajian Bahasa*, *7*(2), 229-246. https://doi.org/10.26499/rnh.v7i2.573
[63] Zalmansyah, A., Hastuti, H. B. P., Saptarini, T., & Budihastuti, E. (2023). The Cultural Identity of Minangkabau and Dayak Kanayatn: An Anthropolinguistic Study. *Eurasian Journal of Applied Linguistics*, *9*(2), 151-162. https://ejal.info/menuscript/index.php/ejal/article/view/560 (taken d/d December 22, 2023)

Sri Kusuma Winahyu, S.S., M.Hum. was born in Yogyakarta, in 1975. In 1997 she completed her undergraduate studies from Gadjah Mada University, Faculty of Cultural Sciences, Department of Indonesian Literature. In 2009 she continued her master degree in University of Indonesia, with a focus on Theoretical Linguistics studies and graduated in 2011. From 2005-2021 she worked at the Language Development and Cultivation Agency and joined the Indonesian Language Proficiency Test Team (*Uji Kemahiran Berbahasa Indonesia*-UKBI) for ten years. In 2024 she completed her doctoral studies in Applied Linguistics from Jakarta State University. Now she works as a researcher at the National Research and Innovation Agency. She focuses on the field of language learning, but is also interested in research on discourse studies. Her Scopus Id: 57216493371; ORCID iD: 0009-0002-9869-6386. Email: sriwinahyu0406@gmail.com or srik004@brin.go.id

Fawwaz Zaini Ahmad, S.Kom. was born in Yogyakarta, Indonesia, in 2000. He is currently as Data Engineer at Bank Rakyat Indonesia, working within the Enterprise Data Management Division. Fawwaz earned his Bachelor's degree in Information Systems from the Sepuluh Nopember Institute of Technology (ITS). His professional focus lies in Data Warehouse Development and Operations, where he specializes in designing, building, and maintaining data warehouse systems to support business intelligence and analytics. He can be reached at fawwaz.zaini@gmail.com

Achril Zalmansyah, M.Pd. has been serving as a researcher at Research Center for Language, Literature, and Community, National Research and Innovation Agency (BRIN), Republic of Indonesia. His works focus on language, literature, culture, and education. He still teaches the Indonesian language at the University of Lampung. Currently, he is continuing his doctoral degree in education at Lampung University. His scholarly works, both nationally and internationally, have been published in journals, proceedings, and book chapters. Email: zzalmansa@gmail.com; ORCID iD: https://orcid.org/0000-0002-3883-5463; Scopus id: 57571704500; and WOS id: AGZ-5062-2022

Exti Budihastuti, S.Pd., M.Pd. was born in Jakarta, Indonesia, in 1966. She is a researcher at the *Research Center for Language, Literature, and Communities-National Research and Innovation Agency (BRIN), Indonesia;* Field of expertise: Applied Linguistics. Meanwhile, the research pursued is about teaching Indonesian for Indonesians, teaching Indonesian for foreigners, and Indonesian literature. She can be contacted at: extibudihasuti@gmail.com or exti001@brin.go.id; Scopus Id: 58523567100; ORCID iD: 0009-0003-9227-7057

Pradicta Nurhuda was born in Pasuruan Regency, June 10, 1991. He is the first researcher at the National Research and Innovation Agency, Republic of Indonesia. She graduated with a Bachelor of Indonesian Language and Literature Education from the Faculty of Letters, State University of Malang, Indonesia (2014). He graduated with a Master of Applied Linguistics from the postgraduate Faculty of State University of Jakarta, Indonesia (2023). His research interest in language, literature, and applied linguistics. With this experience, Pradicta Nurhuda is open to international collaboration and teamwork through prad009@brin.go.id. His Scopus id is: 59137087500; ORCID iD: 0000-0003-1962-3104

Fairul Zabadi was born on February 17, 1965, in Payakumbuh, West Sumatra. He is a researcher at the Research Center for Language, Literature, and Community, National Research and Innovation Agency (BRIN), Indonesia. His research interests are in language, literature, and culture; particularly in learning and education; ongoing research related to local wisdom in forest cultural traditions in Tobelo etnict, in Halmahera, North Maluku; the culture of the Tidung tribe in North Kalimantan. Some works can be seen at https://orcid.org/0000-0002-8521-2277. His email is fzabadi1702@gmail.com

Zainal Abidin, S.S., M.Pd. has his master's degree in language education from Lampung University. He also has expertise in language preservation, especially the scientific and research fields related to language and literature documentation, corpus, and lexicography. Several of his works have been published in both national and international journals. Currently, he works as a researcher at Research Center for Preservation of Language and Literature, National Research and Innovation Agency (BRIN), Republic of Indonesia. Email: zainalwong@gmail.com; zain008@brin.go.id; Scopus Id: 57526481900

Suyadi, S.Pd., M.Si. has been serving as a researcher at Research Center for Language, Literature, and Community, National Research and Innovation Agency (BRIN), Republic of Indonesia. He obtained the title of Master of Science in Social Anthropology. His research interest in language, literature, and education (semiotics, sociolinguistics, linguistic anthropology, and language learning) and his works has been published both in national and international journals/proceedings. Email: suya014@brin.go.id

Sri Yono, S.S., M.Si. was born on June 10, 1970, in Klaten, Central Java, Indonesia. He holds a Bachelor's degree in English Literature from Diponegoro University, Semarang, and a Master's degree in Anthropology from Cenderawasih University, Jayapura, Papua. Currently. He works as a Junior Expert Researcher in Literature at the National Research and Innovation Agency. His writing interests focus on language, literature, and culture. His scholarly works, both nationally and internationally, have been published in journals, proceedings, and book chapters. For detail information please send your email to: sriy009@brin.go.id

Evi Maha Kastri, M.Pd. was born in Tanjung Karang, Lampung, in 1979. She works as a researcher at Research Center for Preservation of Language and Literature, National Research and Innovation Agency (BRIN), Republic of Indonesia. Her career as a researcher began at Lampung Provincial Language Office, Ministry of Education, Culture, Research, and Technology. Her master's education was taken at Master of Indonesian Language and Literature Education, Faculty of Teacher Training and Education, University of Lampung, graduating in 2016. Currently, her functional position is Middle Researcher, her research interests are interdisciplinary linguistics. Email: mahakastri2@gmail.com or evi.maha.kastri@brin.go.id; Scopus Id: 57204046947; ORCID iD: 0000-0003-1305-3593