Accredited SINTA 2 Ranking
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online at: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 9 No. 2 (2025) 283 - 296 e-ISSN: 2580-0760

Word2Vec Approaches in Classifying Schizophrenia Through Speech Pattern

Putri Alysia Azis[1*], Andi Tenriola[2], Dewi Fatmarani Surianto[3], Nur Azizah Eka Budiarti[4], Andi Akram Nur Risal[5], Zulhajji[6]
[1, 2, 3, 4 & 5]Department of Informatics and Computer Engineering, Universitas Negeri Makassar, Indonesia
[6]Department of Electrical Engineering Education, Universitas Negeri Makassar, Indonesia
[1]putrialysia.djarre@gmail.com, [2]andtnriola@gmail.com, [3*]dewifatmaranis@unm.ac.id, [4]nurazizaheka@gmail.com, [5]andiakram@unm.ac.id, [6]zulhajji@unm.ac.id

Abstract

Schizophrenia is a chronic brain disorder characterized by symptoms such as delusions, hallucinations, and disorganized speech, posing significant challenges for accurate diagnosis. This research investigates an innovative Natural Language Processing (NLP) framework for classifying the speech patterns of schizophrenia patients using Word2Vec, with the aim of determining whether there are significant differences between the two features. The dataset comprises speech transcriptions from 121 schizophrenia patients and 121 non-schizophrenia participants collected through structured interviews. This study compares two Word2Vec architectures, Continuous Bag-of-Words (CBOW) and Skip-Gram (SG), to determine their effectiveness in classifying schizophrenia speech patterns. The results indicate that the SG architecture, with hyperparameter tuning, produces more detailed word representations, particularly for low-frequency words. This approach yields more accurate classification results, achieving an F1-score of 93.81%. These results emphasize the effectiveness of the framework in handling structured and abstract linguistic patterns. By utilizing the advantages of both static and contextual embedding, this approach offers significant potential for clinical applications, providing a reliable tool for improving schizophrenia diagnosis through automated speech analysis.

Keywords: *Natural Language Processing, Schizophrenia, Speech Pattern, Word2Vec*

How to Cite: P. A. Azis, T. Andi, D. F. Surianto, N. A. E. Budiarti, A. A. N. Risal, and Z. Zulhajji, “Word2Vec Approaches in Classifying Schizophrenia Through Speech Pattern”, *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 9, no. 2, pp. 283 - 295, Mar. 2025.
DOI: https://doi.org/10.29207/resti.v9i2.6323

1. Introduction

After the COVID-19 pandemic, mental disorders have become one of the problems faced by Indonesian society [1]. Mental disorders are conditions that affect cognition, the ability to manage emotions, or behavior that indicate dysfunction in the psychological, biological, or developmental processes underlying mental function [2]. The DSM-V lists various types of mental disorders, one of which is the spectrum of schizophrenia and other psychotic disorders, which includes schizophrenia, psychotic disorders, and schizotypal disorders. Among these disorders, schizophrenia is the most common in late adolescence to young adulthood, especially in the second and third decades of life. The diagnosis process tends to be complex, especially at certain periods, requiring a deep understanding of its characteristics and its impact on society [3], [4].

Schizophrenia (SZ) is a chronic brain disorder affecting approximately 1.7 per 1000 people or approximately 400,000 people in Indonesia, with symptoms such as delusions, hallucinations, disorganized speech, difficulty in thinking, and low motivation [2], [5]. Globally, the diagnosis of schizophrenia commonly used DSM-IV criteria, which were applied in 47% (N = 7 of 15) of studies. To measure symptom severity, the most frequently used methods were positive and negative syndromic scales, which were applied in 73% (N = 11 of 15) of studies [6].

Research suggests [7], that the mechanisms governing the onset, relapse, symptoms and treatment of schizophrenia (SZ) are still poorly understood, due to a

Received: 21-01-2025 | Accepted: 14-03-2025 | Published Online: 26-03-2025

lack of analytical tools capable of handling the complexity of the disorder. Although the diagnosis of SZ has progressed, the complexity of the disorder often hinders a deeper understanding and the development of effective therapies, calling for more sophisticated analytical solutions, such as artificial intelligence. Recent research has revealed that deep learning, as a branch of artificial intelligence, has the potential to be effective in analyzing the complexities of schizophrenia. One application is to analyze large-scale genomic data and convert genetic variants into images, which can then be classified using CNN algorithms [7], [8]. In addition, other approaches in artificial intelligence, such as Natural Language Processing (NLP), are also potentially effective for analyzing SZ features, especially in analyzing disordered speech.

NLP involves the use of machines to represent and analyze human language computationally, from both phonological and semantic aspects [9], [10]. This research in classifying SZ sufferers based on speech pattern text uses the Word2Vec architecture approach, which is a word representation method consisting of two main models: Continuous Bag-of-Words (CBOW) and Skip-Gram (SG). Both help in generating word representations, with the difference that CBOW predicts words based on their surrounding context, while SG uses the target word to predict its context [11].

Muhammad et al. [12] implemented a Long Short-Term Memory (LSTM) model combined with Word2Vec architecture to process 2,500 review texts from an Online Travel Agent platform. Their approach involved optimizing the model performance by adjusting parameters such as Word2Vec vector dimension, evaluation method, pooling technique, dropout value, and learning rate, which resulted in an accuracy of 85.96%. Meanwhile, Xia. [13] experiment compares the effects of Skip-Gram and CBOW models in word vector training and text classification. The results show that the Skip-Gram model has no significant advantage over CBOW in the classification task. In addition, the use of hierarchical SoftMax and negative sampling showed roughly equivalent performance in this aspect.

Jayadianti et al. [14] emphasized the importance of Word2Vec's CBOW and Skip-gram architectures in enriching datasets with numerical representations that encapsulate semantic nuances, which is crucial for preparing data for subsequent stages of sentiment classification using advanced deep learning methods. Then, Al-Saqqa et al. [15] This research compares the performance of two word2vec models, namely Continuous Bag of Words (CBOW) and Skip-Gram (SG), using Arabic datasets collected from various social media platforms such as Facebook, Twitter, YouTube, and Instagram. The aim of this research is to improve the performance of the detection methods on Arabic. The results show that the CBOW model has higher accuracy than SG, with 74% accuracy for SVM classifier and 72% for RF classifier.

Meanhwile, C. Perira et al. [16] examined the use of Natural Language Processing (NLP) tools to automatically extract labels from radiology reports. This method significantly reduced annotation effort compared to manual labeling. Their research categorized label extraction techniques into four types: Symbolic NLP, Statistical NLP, Neural NLP, and hybrid systems that combine or compare two or more of these approaches.

Recent studies have highlighted the efficacy of these models in mental health applications. For example Voppel et al. [17] using the Word2Vec method to calculate acoustic, and semantic words, and both of the two domain similarity in the classification of patients with schizophrenia spectrum disorders using the random forest method achieved 81% accuracy in acoustic analysis, 80% accuracy in semantic analysis, and 85% from the two domains analysis. In addition, Tsiwah et al. [18] mentioned that schizophrenia spectrum disorder (SSD) and people with Wernice aphasia show semantic deficits in spontaneous speech. This study used the Word2Vec model as a machine learning classification feature to distinguish between spontaneous speech of the two groups, resulting in 81% accuracy.

TaghiBeyglou et al. [19] showed the potential of NLP in detecting Alzheimer's and dementia, especially by analyzing speech pattern disorders in patients. The study showed that the simple architecture used was able to achieve an accuracy of 92% in classifying Alzheimer's cases, and produced a root mean square error of 4.21 to predict Mini-Mental Status Examination (MMSE) scores.

This is particularly relevant in the classification of schizophrenia speech patterns, where understanding the subtleties of language can provide insights into the cognitive and emotional states of individuals. Although the classification of speech patterns of SZ sufferers based on NLP has developed, there are still few studies that discuss in more depth the significant differences in the use of Word2Vec architecture. While various studies have explored NLP-based approaches for analyzing linguistic characteristics in SZ sufferers, limited attention has been given to the comparative effectiveness of Skip-Gram (SG) and Continuous Bag-of-Words (CBOW) in capturing meaningful language patterns, particularly in non-English datasets.

This study aims to bridge this gap by systematically analyzing the performance of SG and CBOW in modeling and classifying speech patterns in SZ sufferers using Indonesian-language datasets. By incorporating data augmentation from Wikipedia and optimizing the Word2Vec hyperparameters, this research evaluates how different embedding strategies influence classification outcomes. The findings are expected to offer new insights into the selection of word embedding techniques for linguistic analysis in schizophrenia research, contributing to the development of more effective diagnostic tools and therapeutic approaches in the field of mental health informatics.

2. Research Methods

This section focuses on the method performed in the research on classification of speech text patterns of people with SZ. The application of the method begins with the collection of SZ and non-SZ datasets. The next step is feature extraction, then text pre-processing to prepare the text in the data training process, then architecture training using Word2Vec, finally evaluation and validation. More details can be seen in the following explanation points presented in visual form in Figure 1.

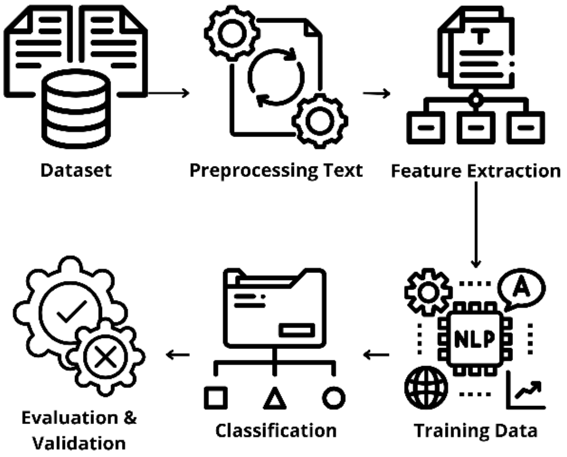

Figure 1. Flowchart Method

2.1 Dataset

Data were collected through in-person interviews at Dadi Regional Special Hospital (RSKD), Makassar City, South Sulawesi, Indonesia. Participation in this study included individuals diagnosed with SZ as well as individuals who did not suffer from SZ, with each group totaling 121 participants with equal numbers of males and females.

The interview process was conducted using a series of pictures as stimulus as presented in Figure 2. The two types of pictures used were: (a) pictures of simple houses, which aim to trigger structural and emotive descriptions, and (b) Rorschach pictures, which are used to explore deeper psychological responses. Participants were asked to describe what they saw in the pictures, with the aim of collecting data on speech patterns that could indicate cognitive and emotional features associated with SZ or non-SZ conditions.

Due to ethical considerations and confidentiality agreements, the Rorschach images used in the study cannot be publicly disclosed. The inclusion of these images in the study was approved under ethical guidelines, ensuring their usage adhered to research integrity and participant protection. The exclusion of direct visual representation does not affect the validity of the study, as the analysis is based on participants' verbal responses rather than the images themselves.

(a)

(b)

Figure 2. Image test: (Simple house image, sorce: https://pin.it/5DDbgvRlJ and (b) Rorschach image)

The data generated from the interviews was further analyzed to identify certain patterns in speech that could distinguish between SZ and non-SZ sufferers. In the practice of collecting datasets, this research uses data that is balanced between the two classes, this is done to avoid bias during the data training process. In addition, this data balance allows the model to learn the characteristics of both classes proportionally. With balanced data, the evaluation of model performance is also expected to be fairer so that the resulting metrics such as accuracy are not much different between SZ and non-SZ classes.

Furthermore, the dataset obtained in the form of audio is manually transformed into text, resulting in a text dataset in the form of an Excel file. This dataset transformation does not use additional techniques or software, with the hope that this transformation does not change the original meaning of the speech text patterns of SZ and non-SZ sufferers, an example of the resulting dataset can be seen in Table 1.

2.2 Preprocessing Text

One of the important processes in this research is text preprocessing. Preprocessing text serves to help avoid potential errors that interfere with data processing [20]. In this research there are three stages of preprocessing, first manually preprocessing to convert audio data into text form collected in Excel files, where at this stage everyday words are converted into standard words in Indonesian, for example, the word “pohong” to “tree”, but does not change the actual meaning. Furthermore, the data collected in Excel files is given the treatment of techniques in NLP pre-processing, namely case folding and regex. Case folding is a stage to convert letters in the text into upper or lower-case letters [21], while regular expression (regex) is an algebraic notation to specify a set of strings that serves to return all text that matches the patterns in the data corpus [22]. These two pre-processing methods were chosen because they do not remove the essence of the meaning of the participants' speech patterns.

Table 1. Sample dataset

Picture	Sex	Class	
		SZ	Non-SZ
Image (a)	Male	Gambar rumah, ada mataharinya ada awan ada pohon ada taman bunganya, ada pagar, ada terasnya, ini ada batu teras halaman. Ini rumahnya fungsinya untuk ditinggali, matahari ini untuk menyinari dunia, awan ini untuk berlindung atau menurunkan air hujan, pohon sebagai untuk kehidupan, pagar ini untuk halaman teras untuk menutupi rumah.	Terdapat pagar yang berfungsi membatasi dan melindungi rumah di belakangnya. Terdapat susunan batu sebagai jalan setapak menuju pintu rumah. Terdapat pintu yang menjadi jalur keluar dan masuk dengan keadaan tertutup agar keadaan dalam rumah tidak terganggu hal dari luar. Terdapat ventilasi udara. Terdapat atap sebagai pelindung rumah dari atas. Terdapat beberapa semak dan pohon di pekarangan rumah untuk menghijaukan sekitar rumah.
	Female	Ini matahari adalah sumber cahaya yang menerangi dunia, tetapi kadang-kadang juga menjadi pemicu kegelapan dalam pikirannya yang gelap. Awan mengumpulkan air di langit, membentuk hujan yang membasahi tanah, namun juga bisa menjadi awal dari badai pikiran yang mengamuk. Rumah adalah tempat perlindungan dari badai kehidupan, tetapi juga penjara dari kekacauan yang berputar-putar dalam benaknya. Pepohonan memberikan oksigen yang menyegarkan, tetapi juga bisa menjadi saksi bisu atas kekosongan dalam pikirannya. Bunga-bunga menawarkan aroma yang manis, tetapi juga bisa menjadi ilusi dalam khayalan yang terus berputar.	Gambar di atas merupakan sketsa rumah yang dimana terdapat pagar, pohon, semak-semak, dan juga bunga. Dan terlihat bahwa cuaca dalam sketsa cerah. Yang dibuktikan dengan adanya matahari yang digambarkan didalam sketsa tersebut.
Image (b)	Male	Warna-warni, merah, biru, hijau, kuning. Menyerupai warna saja.	Yang saya lihat lukisan abstrak kalau orang awam lihatnya seperti kupu-kupu dikarenakan gambar yang selaras, warnanya ada biru, merah, kuning, hijau, orange, ada merah gelap juga dan coklat.
	Female	Warna biru, warna kuning, warna hijau, dengan warna putih. Dan seperti paru-paru oksigen.	Dari gambar ini saya melihat interpretasi dari beberapa hewan, beberapa hewan itu saya melihat seperti kupu-kupu, laba-laba dan kuda laut.

2.3. *Feature Extraction*

Feature extraction is the process of selecting and transforming data into digital features that can be processed while maintaining the information contained in the original text. In this research, the feature used is Word2Vec by utilizing both Skip Gram (SG) and Continuous Bag-of-Words (CBOW) methods.

The selection of the Word2Vec method in this research is due to the use of a relatively small dataset, besides that, this research focuses on semantic relationships or data with high context variations. The use of both features in Word2Vec is used to see if the two features have differences in the same data.

Research indicates that while Word2Vec effectively captures semantic relationships between words, its performance is significantly affected by the size of the training dataset. Studies have shown that smaller datasets can lead to suboptimal embeddings, as the model may not sufficiently learn the contextual nuances necessary for accurate classification tasks [23]. Specifically, CBOW is often recommended for smaller datasets due to its efficiency in predicting words based on their context. However, it still faces challenges in fully capturing the complexity of language [23].

Moreover, recent findings emphasize that the effectiveness of Word2Vec diminishes when applied to datasets that do not provide sufficient examples for the model to learn from, which is particularly relevant in clinical settings where data availability is limited [24]. For instance, a study highlighted that relying on external word pairs for hyperparameter tuning is suboptimal for smaller datasets, suggesting that the inherent limitations of Word2Vec become more pronounced in data-scarce environments [24]. This aligns with the broader consensus in NLP research that larger datasets generally yield better performance, as they allow for more robust model training [25]. Thus, while Word2Vec remains a valuable method, its application to a small dataset, as in this study, is supported by prior research demonstrating that meaningful results can still be obtained with careful selection of hyperparameters and validation methods.

SG architecture is an architecture that tries to maximize the classification of a word based on other words in the same sentence. More specifically, each of these words is used as input for a log-linear classifier with a continuous projection layer and predicts words in a certain range before and after the word [26]. The mathematical architecture can be seen in Formula 1.

$$Q = C \times (D + D \times \log_2(V)) \quad (1)$$

Where C is the maximum distance between words. For example, if C = 5, then for each word, the training will randomly select a number R in the range <1; C>. CBOW architecture is an architecture that removes the non-linear hidden layer and the projection layer is shared for all words so that all words can be projected to the same position where the word vector is shared equally so that it can be called Continuous Bag-of-Words because the order of words in context does not affect the projection. The mathematical architecture can be seen in Formula 2.

$$Q = N \times D + D \times \log_2(V) \quad (2)$$

2.4 Training Data

After feature extraction is complete, the next step is to utilize the generated features to train the machine learning model. The training process aims to build a model that can recognize patterns and relationships in the data based on the vector representation generated by the Word2Vec method. At this stage, the Skip-Gram (SG) and Continuous Bag-of-Words (CBOW) algorithms are used to explore the influence of each approach on the quality of the embedding formed. By utilizing these two algorithms, this research attempts to understand how broader context (captured by SG) and local context (which is more focused on words around the target as in CBOW) affect word representation and, ultimately, model performance in classification tasks.

This training stage also involves optimizing the convergence process, which depends on the choice of regularization technique, the number of iterations, and the batch size used. During training, the model will gradually learn to associate words in a more semantic context and build a more accurate representation of the relationships between words in the text. Therefore, the results of training depend heavily on the way the feature extraction method works in describing the information contained in the data corpus. In other words, the training process not only creates a model that can recognize patterns in the data but also reflects the effectiveness and reliability of the feature extraction method applied

The stage process is used to prepare the dataset at the data training stage. The data training stage in this research is divided into 2, based on the use of the Word2Vec method features, namely SG and CBOW and will be explained in more detail.

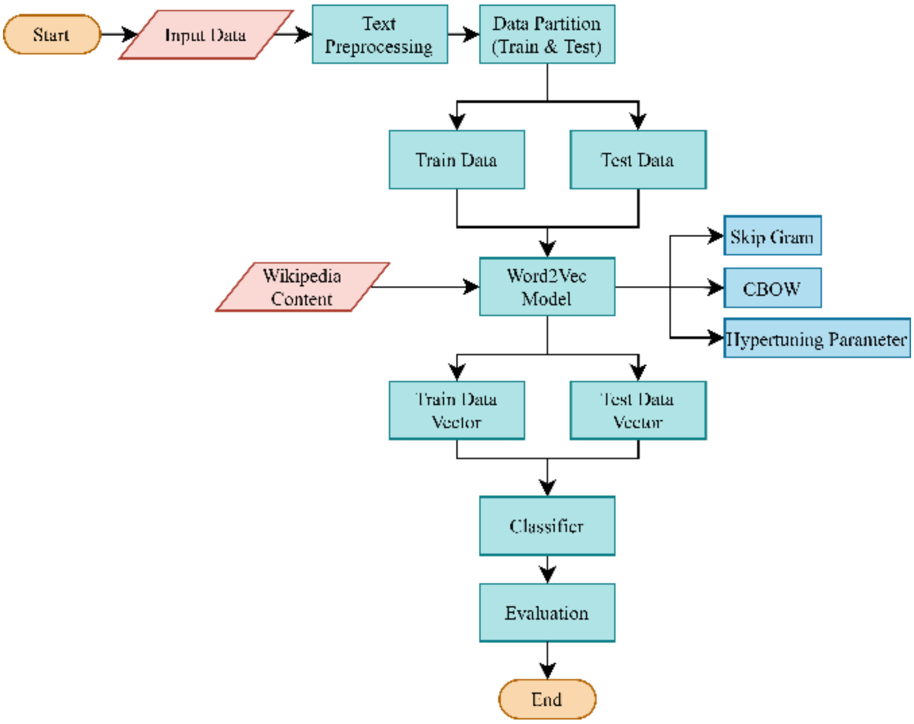

Figure 3. Training Data Flow Diagram

It can be seen in Figure 3 that the process flow at the data training stage includes several main stages. The process starts with the input of Excel data that will be used as input. After that, the data goes through a text preprocessing stage which includes case folding and regex. After that, the data is divided into two partitions, namely training data (80%) and test data (20%). The processed corpus is then enriched with Wikipedia content, which is added to create a more comprehensive embedding when power training begins.

The next stage is the formation of the Word2Vec model, where the Skip Gram (SG) and Continuous Bag of Words (CBOW) algorithms are used to generate a vector representation of the words in the corpus. This process also involves hyperparameter tuning to obtain an optimal model. Next, the Word2Vec result vector is then split into two, namely train data vector and test data vector. The data is then classified using Random Forest and SVM methods, according to the experiments conducted, to test the performance of the classification model. The process ends with the evaluation stage, where the performance of the model is evaluated according to relevant performance metrics to determine the effectiveness of the methods performed in this study.

2.5 Classification

The next research method involves the classification process using the random forest and support vector machine algorithm. Random forest is an ensemble-based classification method that consists of a set of classifiers in the form of a decision tree {h (x, Θk), k = 1, ...}, where {Θk} is an identically distributed and independent random vector. Each tree in this model casts one vote to determine the most dominant class based on input x [27].

Random Forest is a robust ensemble learning method that operates by constructing multiple decision trees during training and outputting the mode of the classes for classification tasks. One of the primary reasons for selecting RF is its ability to handle high-dimensional data, which is often encountered in mental health analysis. Yu et al. found that RF outperformed other machine learning algorithms in discriminating schizophrenia patients from healthy controls, achieving an area under the curve (AUC) of 0.886, which highlights its effectiveness in this domain [28]. The ensemble nature of RF allows it to mitigate overfitting, a common issue when working with smaller datasets, by averaging the predictions of multiple trees, thereby enhancing generalization.

Moreover, RF's capability to provide insights into feature importance is particularly beneficial in the context of schizophrenia classification. The algorithm can rank the significance of different features, allowing researchers to identify which speech patterns or neuroimaging markers are most indicative of schizophrenia. This feature selection capability is critical for understanding the underlying characteristics of the disorder and for refining diagnostic criteria. The study by Gashkarimov et al. emphasizes that RF can effectively identify relevant features in complex datasets, making it a valuable tool for psychiatric research [29].

On the other hand, Support Vector Machines (SVM) are particularly effective for classification tasks involving high-dimensional spaces. SVM works by finding the optimal hyperplane that maximizes the margin between different classes, making it highly effective for binary classification problems. Jo et al. noted that SVM has been widely used in schizophrenia studies due to its ability to classify different classes effectively, even with limited training data [30]. This characteristic is especially relevant when working with smaller datasets, as SVM can still achieve high accuracy by focusing on the most informative features.

Moreover, Support Vector Machine (SVM) method is a machine learning algorithm used to solve classification and regression problems by maximizing the margin between data classes in the feature space. SVM is also famous for its ability to handle non-linear data through the use of kernel functions. Where this method has the main idea, viz: non-linear mapping to a high-dimensional space, hyperplane with maximized margins, and high generalization ability [31].

These two methods were chosen to classify SZ and non-SZ using both architectural methods in Word2Vec, this was done to see if there was a significant difference based on the classification method between the two methods. In addition, these two methods were chosen based on their respective advantages such as random forest which has flexibility that allows this method to work optimally with feature representations, and SVM which has flexibility with kernel functions that can transform data to a high-dimensional feature space without explicitly calculating the transformation.

2.6 Evaluation and Validation

After going through the classification process, the next step is to validate and evaluate the performance of the model architecture. This stage aims to ascertain whether there is a significant difference in results between the two architectures being compared. In the validation and evaluation process, the confusion matrix is used as the main tool to provide a deeper understanding of the model performance in a visual form, especially in the context of classification. The confusion matrix not only shows the distribution of correct and incorrect predictions, but also provides insight into the types of errors that occur, such as false positives and false negatives. This stage is very important to determine the advantages and disadvantages of each approach used.

Precision describes how precise the architecture is in predicting the positive class, while recall indicates how well the architecture recognizes instances of the positive class. Meanwhile, accuracy measures how accurate the architecture is in predicting all classes. The values in the Confusion Matrix are usually expressed in percentage (%) [32], which gives a clear picture of the architecture's prediction quality. The mathematical form can be seen in Equations 3 – 5.

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \times 100\% \quad (3)$$

$$Precission = \frac{TP}{TP+FP} \times 100\% \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \tag{5}$$

Then, the comparison between the average precision and recall values obtained is called the F-1 Score, which can then be explained in its mathematical form can be seen in Equations 6.

$$F1\ Score = \frac{2\ X\ Precision\ X\ Recall}{Precision\ X\ Recall} \tag{6}$$

TP (True Positive) refers to the number of texts that are truly SZ and correctly identified as SZ by the model. TN (True Negative) refers to the number of texts that are truly non-SZ and correctly identified as non-SZ by the model. FP (False Positive) refers to the number of texts that are actually non-SZ but are incorrectly identified by the model as SZ, meaning that the model makes a mistake by assigning a positive label to data that should be negative. Whereas FN (False Negative) refers to the number of texts that should have been identified as SZ, the model misclassified them as non-SZ, ignoring the fact that they are actually related to schizophrenia disorder. This measure is important for evaluating the model's performance in classifying data, as it provides a clear picture of the extent to which the model is able to correctly identify positive and negative cases, as well as the potential errors that occur in classifying data related to these conditions.

To clarify the accuracy results in this study, another classification metric in machine learning is also used, namely the Receiver Operating Characteristic (ROC) Curve. This metric provides more comprehensive information about the performance of the classification model compared to traditional metrics such as accuracy. The ROC Curve illustrates the trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR) at different thresholds. In addition, the Area Under the Curve (AUC) measures the area under the ROC Curve, which provides a numerical score to assess the model's ability to distinguish between positive and negative classes. A higher AUC value indicates that the model is more effective in distinguishing the two classes [33]. The mathematical form for TPR and FPR can be seen in Equations 7 – 8.

$$True\ Positive\ Rate = \frac{TP}{TP+FN} \tag{7}$$

$$False\ Positive\ Rate = \frac{FP}{FP+TN} \tag{8}$$

The AUC is calculated as the area under the ROC curve. One way to estimate it is through the trapezoidal rule, which integrates the points on the ROC curve. For two consecutive points on the curve (FPR_i, TPR_i) and (FPR_{i+1}, TPR_{i+1}), the area under the segment is computed as Equation 9.

$$Area_i \frac{(FPR_{i+1} - FPR_i)\ .\ (TPR_{i+1} + TPR_i)}{2} \tag{9}$$

The total AUC is the sum of the areas for all segments between consecutive Equation 10.

$$AUC = \sum_i Area_i \tag{10}$$

Furthermore, the performance metrics in this study use K-Fold Cross-Validation to measure the predictive ability of machine learning models more fairly. This technique helps minimize the bias that can occur when relying on only one division of data for training and testing. The process starts by dividing the data into k subsets (folds). The model is then trained using k-1 folds as the training set, while the remaining 1 fold is used as the test set. This process is repeated k times, so that every subset of data has been used as a test set. After all iterations are complete, the evaluation results from each trial are averaged to provide an estimate of the overall model performance [34], the mathematical formula can be seen in Equation 11.

$$M_{CV} = \frac{1}{k}\sum_{i=1}^{k} M(f_{\theta(i)}, D_i \tag{11}$$

This study also considers computation time as an additional evaluation metric. Computation time is a crucial performance indicator in algorithm analysis, which is used by software scientists to assess the execution efficiency of an algorithm as well as to ensure that the developed model can operate within a reasonable and optimal time range [35]. The decision of computation time as an evaluation metric is based on several key considerations, namely that in real-world applications, an accurate model that requires a long computation time is a limitation of a model. In addition, the efficiency of execution time is also related to the fact that models that have high performance but low computation time are easier to apply to larger datasets. Thus, evaluation based on computation time not only measures the effectiveness of the model in terms of accuracy but also in terms of efficiency and applicability in various usage scenarios.

3. Results and Discussions

This study aims to detect speech patterns as one of the main symptoms of Schizophrenia (SZ) by utilizing the Word2Vec method. Word2Vec is used to represent speech patterns in vector form, thus enabling in-depth analysis of the semantic and syntactic relationships typical of SZ sufferers. This study compares the performance of two key features in Word2Vec, namely Continuous Bag of Words (CBOW) and Skip Gram (SG), to evaluate how each feature affects classification accuracy.

In measuring the effectiveness of these approaches, classification performance is evaluated using various metrics, such as accuracy, precision, recall, and F1 score. This analysis aims to determine the significant differences between the two approaches, as well as provide further insight into the potential of Word2Vec in supporting data-driven SZ diagnoses. In practice, this research utilizes 3 experimental approaches, which will be explained further.

3.1 Default Parameter

In the initial experiments, this research applied default parameters to the Word2Vec feature, as implemented in the gensim library. Default parameters such as vector size = 100, window = 5, and min count = 5 are often used to produce efficient word embedding. This approach was used to test Word2Vec's ability to detect speech patterns in people with Schizophrenia (SZ), as Word2Vec was also applied in previous research for word embedding-based tasks with default [36].

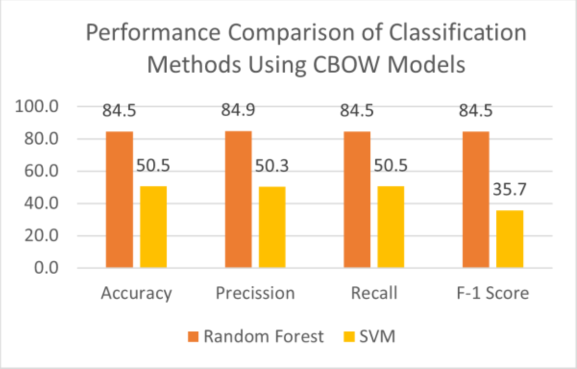

Figure 4. Performance Metrics using CBOW

Figure 4 shows the performance results of the Continuous Bag of Words (CBOW) architecture using two different classification methods. Based on the data, the Random Forest classification method shows superior performance with accuracy, precision, recall, and F1 value of 84%, while the SVM classification method only managed to achieve accuracy, precision, and recall of 50%, and F1 of 35,7%. This difference in performance is caused by various factors, one of them being the characteristics of each classification method on the data used.

Based on its characteristics, Random Forest (RF) uses the bootstrap aggregating (bagging) technique, which is a machine learning method that combines multiple models to improve the accuracy and stability of the algorithm. This technique allows RF to perform well on datasets with a high level of complexity. In contrast, SVM has the advantage of handling high-dimensional data with non-linear kernels. However, on datasets with highly complex distributions or difficult to optimally separate by a hyperplane, SVM performance may suffer.

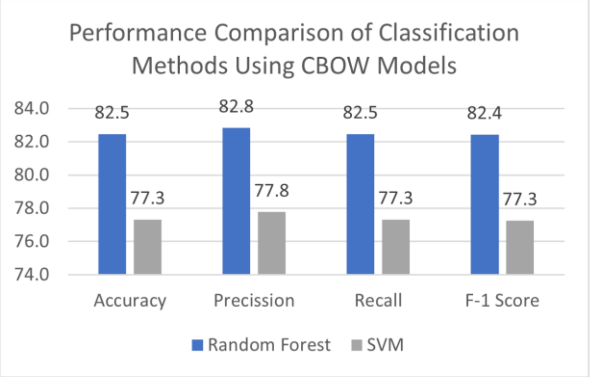

Figure 5. Performance Metrics using SG

Figure 5 presents the performance results of the Skip Gram (SG) architecture with two different classification methods. Based on the data, the Random Forest classification method shows superior performance with accuracy, precision, recall, and F1 of 82%. While the SVM method only managed to achieve accuracy, recall, F1 of 77,3% and precision of 77,8%.

Figures 4 and 5 present a comparison of classification methods used in this study. Both Word2Vec architectures, Continuous Bag-of-Words (CBOW) and Skip-Gram (SG) performed better with the Random Forest (RF) classification method. The RF classification method shows superiority in classifying the speech pattern dataset of people with schizophrenia (SZ). This advantage is supported by RF characteristics that require minimal preprocessing and are relatively resistant to differences in feature scale. In contrast, SVM is sensitive to data scale and requires more careful preprocessing. In this dataset, the data was not given in-depth preprocessing due to its nature which tends to change meaning if too many modifications are made.

The Schizophrenia (SZ) speech pattern dataset also contains pattern features that reflect interactions between words and sentences. Random Forest (RF) excels in capturing interactions between features due to its capability of aggregating results from multiple decision trees. In contrast, Support Vector Machine (SVM) primarily focuses on determining the optimal decision boundary for class separation, which makes it less effective in modeling complex feature relationships within the dataset.

3.2 Additional Data Train

The results of the first experiment are shown in Figure 4 and Figure 5, the Random Forest (RF) classification method has an advantage over the Support Vector Machine (SVM) based on the dataset used. Based on these findings, the second experiment focused solely on using the RF classification method, due to its better ability to capture the meaning of the dataset used in this study.

In the second experimental setup, the training dataset to build the Word2Vec model was expanded by incorporating additional relevant content from Wikipedia using the Wikipedia API. This approach aimed to enrich the word representations in the word2vec model by integrating domain-specific textual data related to schizophrenia.

The process involved selecting a set of relevant topics encompassing schizophrenia, mental disorders, treatment methods, general mental health, and technological applications in psychiatric research. These topics were then used to extract textual content from Wikipedia through the Wikipedia API. The API function verified the existence of each topic’s page and retrieved the corresponding text. The collected text was then preprocessed and added to the training corpus of the word2vec model. By integrating this enriched dataset, the model was expected to learn more comprehensive word embeddings, capturing deeper semantic relationships and improving its contextual understanding of schizophrenia-related terms.

This addition aims to enrich the vocabulary, so as to understand more words and phrases and produce a more diverse and in-depth representation. In addition, the use of Wikipedia articles is expected to improve the quality of semantic representations due to their descriptive nature and include interconnections between concepts. Thus, the vector representation generated by Word2Vec is expected to be more effective in capturing the semantic meaning of words.

To evaluate the impact of additional training data from Wikipedia, the word2vec model was trained using two different architectures: Continuous Bag of Words (CBOW) and Skip-Gram (SG). The evaluation focuses on key performance metrics, including accuracy, precision, recall, F1-score, and training time as presented in Table 2.

Table 2. Percentage comparison of SG and CBOW + Wikipedia

Model	Accuracy	Precision	Recall	F1	Time (S)
CBOW	76.29	76.4	76.29	76.27	0.16
SG	79.38	80.24	79.38	79.26	0.34

According to Table 2, SG shows superior performance with accuracy, recall, F1 of 79% and precision of 80%, compared to CBOW which reaches 76% each accuracy, recall, F1 and precision. However, in terms of training time, CBOW is more efficient with a duration of only 0.16 seconds, while SG takes 0.34 seconds. These findings suggest that CBOW remains a suitable choice when computational efficiency is a priority, whereas Skip-Gram is more effective for capturing complex word relationships, making it preferable for tasks requiring higher semantic accuracy.

Figure 6 below illustrates the performance comparison of the CBOW model trained on the original dataset versus the dataset augmented with additional text from Wikipedia. The evaluation is based on four key metrics: accuracy, precision, recall, and F1-score.

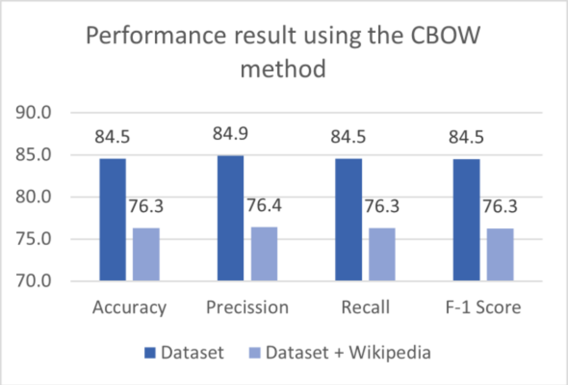

Figure 6. Performance Metrics using CBOW between experimental 1 and 2

Based on Figure 6, the comparison between experiment 1 and experiment 2 using additional data from Wikipedia shows that the addition of data from Wikipedia to the training data causes a significant decrease in the performance of the CBOW model in capturing the semantics of words in the dataset. This decrease can be seen in the accuracy, precision, recall, and F1-Score which are all at 84% on the model trained only with the dataset, while on the dataset that added Wikipedia data is at 76%.

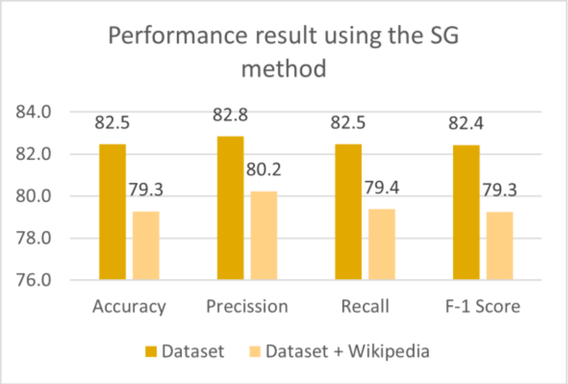

Figure 7. Performance Metrics using SG between experimental 1 and 2

As figure 7 is a method that uses the Skip Gram (SG) approach, the analysis results show that the addition of data from Wikipedia to the training data actually causes a slight decrease in the model performance metrics. This decrease can be seen in the accuracy, recall, F1-score at 79% and precision at 80% after the Wikipedia data is added. In comparison, the model trained using only the main dataset without additional Wikipedia data was able to achieve a higher performance metric of 82%.

Based on both figures (Figure 6 and Figure 7) the model has a significant performance degradation when matched with the Wikipedia-added dataset. Accuracy, precision, recall, and F1-score are all lower compared to the default dataset model. This indicates that the addition of Wikipedia data may have introduced noise or irrelevant contextual relationships, affecting the quality of the learned word embeddings.

Based on these experimental results, there are several factors that could be responsible. One of the main factors is the characteristics of both Continuous Bag of Words (CBOW) and Skip-Gram (SG) models, which build word representations based on the context in the datasets used for training. The first experiment tends to have a more structured dataset when compared to the second experiment which was retrained with additional Wikipedia data, which tends to be freer although still in the context of schizophrenia (SZ).

Furthermore, when looking at the quality of the dataset used, Wikipedia has a very different writing style to the default dataset. Wikipedia articles often use more formal, technical, or highly technical language which can make it difficult for the model to capture the context or more common words in the default dataset. Whereas the default dataset contains many conversations, which are taken directly from schizophrenia (SZ) or non-SZ patients according to the images shown.

3.3 Hyperparameter Tuning

The last experiment was conducted by applying hyperparameter tuning to the Word2Vec architecture, where parameter settings such as vector size = 300, window = 5, and min count = 2 are frequently used in previous literature. The study shows that the use of a vector size of 300 and other parameter variations, such as window size and min count, can affect the embedding quality, especially in downstream tasks such as named entity recognition (NER) and sentiment analysis [37].

In this study, the main context-rich dataset is still used, but the Wikipedia dataset is added to expand and enrich the representation of word context. The utilization of Wikipedia was chosen due to its information-rich nature and cover a wide range of topics, thus being able to support a more in-depth semantic analysis. A vector size of 300 was chosen to ensure the model's ability to capture semantic details from the combination of the two datasets.

Table 3. Percentage comparison of Hyperparameter Tuning

Model	Accuracy	Precision	Recall	F1	Time (S)
CBOW	89.69	89.97	89.69	89.68	1.18
SG	93.81	93.89	93.81	93.81	3.07

In the third experiment, the impact of applying hyperparameter tuning on the performance of the two Word2Vec architectures is evident in Table 3. These parameter changes significantly improved the performance metrics, where the Continuous Bag of Words (CBOW) architecture achieved accuracy, precision, recall, and an F1 score of 89%. Meanwhile, the Skip-Gram (SG) architecture shows higher performance with metrics that reach 93% for all these measures.

In addition to the performance improvement, the application of these parameters also has an impact on the computation time. Both architectures require higher computation time compared to previous experiments, with SG taking 3.07 seconds longer than CBOW with 1.18 seconds. This is consistent with the more complex characteristics of SG in capturing word representations in sparse contexts. Thus, hyperparameter tuning contributes positively to the performance of both architectures, although it comes with an increased computation time requirement.

Figure 8 presents the confusion matrix for the classification task performed using the CBOW model, which was trained on the original dataset along with additional Wikipedia data and further optimized through word2vec hyperparameter tuning. The classification was conducted using the Random Forest classifier, and the matrix provides insight into the model's predictive performance.

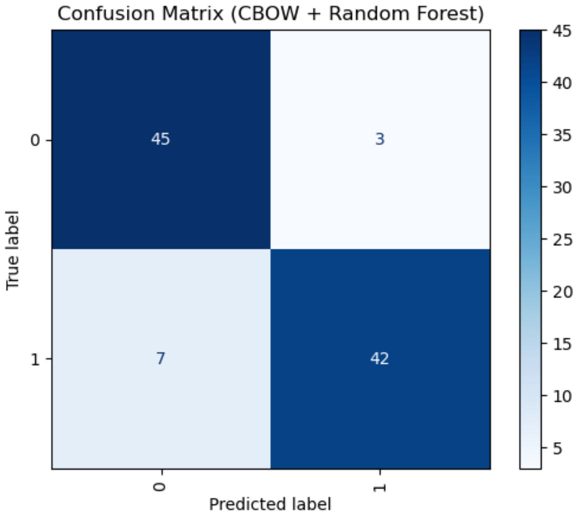

Figure 8. Confusion Matrix CBOW Hypertuning

Based on the matrix, the CBOW model successfully identified the Schizophrenia (SZ) class with a True Positive (TP) of 45. In addition, the model also accurately classified the True Negative (TN) for the non-SZ class of 42.

These results suggest that while the incorporation of Wikipedia data and hyperparameter tuning has improved classification accuracy, some misclassification errors persist. Further refinement, such as additional feature engineering or alternative classification models, could be explored to enhance performance.

Figure 9 presents the confusion matrix for the classification task performed using the Skip-Gram (SG) model, trained on both the original dataset and additional Wikipedia data, with further optimization through word2vec hyperparameter tuning. The classification was conducted using the Random Forest classifier, and the confusion matrix provides insight into the model's predictive performance.

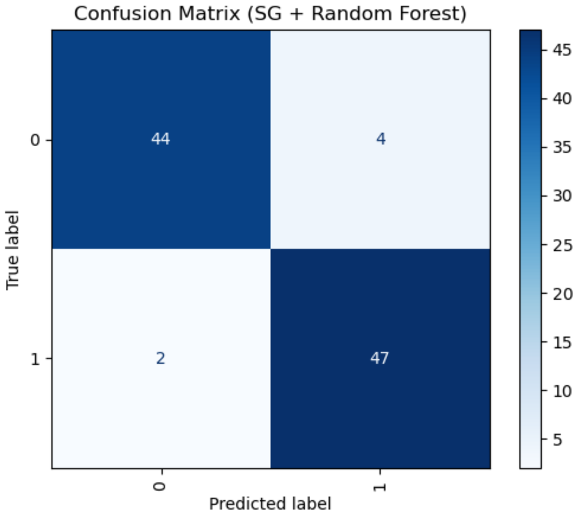

Figure 9. Confusion Matrix SG Hypertuning

Based on the matrix, the SG architecture was able to identify the Schizophrenia (SZ) class with 44 True Positives (TP), indicating the model's success in recognizing SZ class samples with a high degree of accuracy. In addition, the model also successfully classified True Negative (TN) for non-SZ classes as many as 47, confirming the reliability of SG in recognizing samples from non-SZ classes. These results demonstrate that the SG architecture has superior capabilities compared to CBOW in identifying complex patterns in SZ speech text, reinforcing SG's position as a more effective approach for text data-based classification tasks with rich context.

These findings highlight that Skip-Gram, in combination with Random Forest, demonstrates higher robustness and generalization ability in this classification task. The lower misclassification rates indicate that SG benefits more from Wikipedia data augmentation and hyperparameter tuning compared to CBOW. However, further optimization, such as fine-tuning the classifier or incorporating additional linguistic features, could still be explored to further reduce misclassification errors.

Furthermore, the evaluation of performance metrics in this study is also enriched with other evaluation methods, one of which is the Receiver Operating Characteristic (ROC) Curve. This method is used to analyze the model's ability to distinguish between positive and negative classes, thus providing a more comprehensive picture of the model's performance.

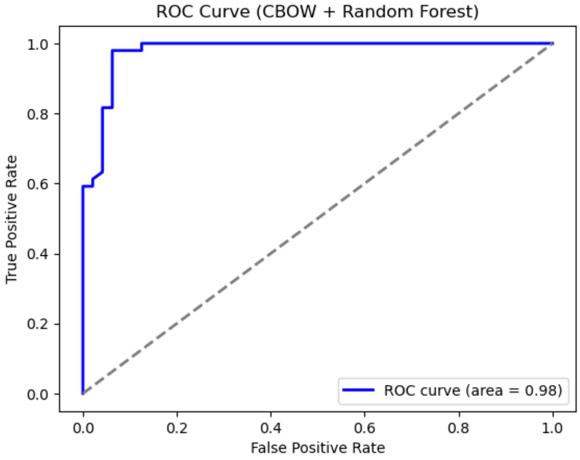

Figure 10. CBOW ROC Curve

In order to clarify the results obtained from the simple metric method in the third experiment, this study also used the Receiver Operating Characteristic (ROC) Curve evaluation method. In Figure 10, the results of applying the ROC metric based on the third experiment using the Continuous Bag of Words (CBOW) method are shown. The evaluation results show an improvement over the previous metrics, with the ROC score reaching 97.85%. This indicates that the model has a better performance in distinguishing between positive and negative classes compared to the previous evaluation method.

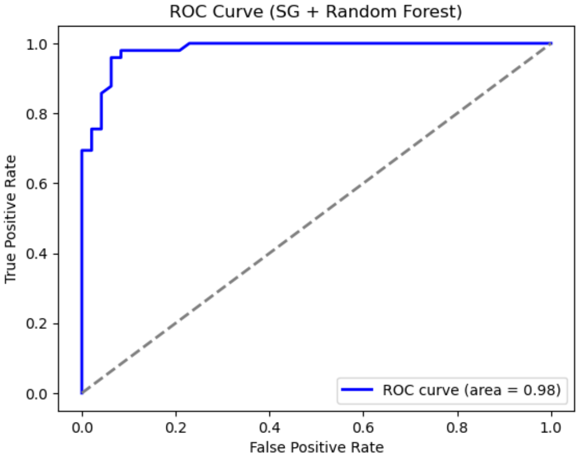

Figure 11. SG ROC Curve

In Figure 11, the evaluation results using the Receiver Operating Characteristic (ROC) Curve with the Skip-Gram method are compared with the previous metrics in evaluating model performance in the third experiment. The results show an improvement, with the ROC score reaching 98.21%, indicating that the model performs better than the previous evaluation.

Evaluation using the Receiver Operating Characteristic (ROC) Curve shows that both the Continuous Bag of Words (CBOW) and Skip-Gram (SG) methods perform well in classifying the speech patterns of schizophrenia (SZ) patients. In addition, the results obtained from these two methods do not exhibit statistically significant differences. To further enhance the robustness of the performance evaluation, this research incorporates Cross-Validation as an additional metric to assess the classification of the dataset. This approach ensures a more reliable and comprehensive analysis of the model's performance by mitigating potential biases and improving generalizability.

Table 4. Percentage comparison with Cross-Validation

Model	Accuracy	Precision	Recall	F1
CBOW	93.3	93.42	93.3	93.29
SG	94.32	94.54	94.32	94.31

Table 4 presents a comparison of performance evaluation using the Cross-Validation method. The results demonstrate an improvement over a simple metric-based evaluation. Notably, Skip-Gram (SG) outperforms Continuous Bag-of-Words (CBOW), achieving a higher accuracy rate of 94%.

The results of the three experiments show that the Word2Vec method is effective in detecting speech patterns in people with Schizophrenia (SZ), with the Skip-Gram (SG) architecture consistently outperforming Continuous Bag of Words (CBOW). In the first experiment, the use of default parameters resulted in a fairly good initial performance, with CBOW accuracy reaching 84% and SG reaching 82%

using the Random Forest (RF) classification method. The second experiment showed that the addition of data from Wikipedia improved the semantic representation, although its effect on model performance decreased with CBOW at 76% and SG at 79%.

The third experiment demonstrated that hyperparameter tuning, with a vector size of 300, a window size of 5, and a minimum count of 2, significantly improved both architectures. The Skip-Gram (SG) model achieved the highest accuracy, recording 93% based on simple metrics, 98.21% on the ROC Curve, and 94% on Cross-Validation. In comparison, the Continuous Bag-of-Words (CBOW) model achieved 89% on simple metrics, 97.85% on the ROC Curve, and 93% on Cross-Validation. These results indicate that SG is more effective in capturing complex linguistic patterns characteristic of schizophrenia (SZ) speech, particularly after parameter optimization.

This research demonstrates key strengths in the use of Word2Vec to detect complex linguistic patterns in people with Schizophrenia (SZ), particularly through its Skip-Gram (SG) architecture that is able to capture semantic relationships in a broader context. These findings highlight the importance of model selection, dataset curation, and parameter optimization in the development of NLP-based diagnostic tools for schizophrenia. However, this study has several limitations, including the model's sensitivity to dataset size. The addition of data from Wikipedia, although enriching the word representation, does not always result in significant performance improvements, especially in the SG architecture which tends to be affected by the distribution of new data. In addition, the computational requirements increased significantly in experiments with hyperparameter tuning, which is an important factor in large-scale applications. Factors such as the dataset distribution, the complexity of the SZ speech pattern, and the level of data preprocessing also affect the results achieved in this study. For future research, testing with larger datasets and the use of additional data enrichment techniques can be conducted to strengthen the generalizability of the architecture.

Based on Table 5, this study shows the best results compared to previous studies in the classification of schizophrenia spectrum disorders. With the Word2Vec + Random Forest (RF) method, this study achieved an accuracy rate of 93.81%, higher than the previous studies that used similar approaches. For example, a 2023 study that combined Word2Vec and Random Forest achieved 85% accuracy, while another study that combined Word2Vec with Word Error Rate only achieved 77%. Thus, this research makes a significant contribution to the detection of schizophrenia spectrum disorders using natural language processing and machine learning, while confirming the effectiveness of the methods used.

Table 5. Comparison of Results with Previous Research

No	Title / Year	Dataset	Method	Result
1	Semantic-based NLP techniques discriminate schizophrenia and Wernicke’s aphasia based on spontaneous speech / 2024	schizophrenia spectrum disorder and Wernick’s aphasia	word2vec and sBERT + Random Forest	81%
2	Context is not key: Detecting Alzheimer’s disease with both classical and transformer-based neural language models / 2024	Alzheimer’s disease	Word2Vec + GPT + BERT	92%
3	Semantic and Acoustic Markers in Schizophrenia-Spectrum Disorders: A Combinatory Machine Learning Approach / 2023	Schizophrenia Spectrum Disorders	Word2Vec + Random Forest	85%
4	Combining automatic speech recognition with semantic natural language processing in schizophrenia / 2023	Schizophrenia Spectrum Disorders	Word2Vec + Word error rate	77%
5	Schizophrenia classification using machine learning on resting-state EEG signal / 2022	Schizophrenia Spectrum Disorders	DeepSeek + Principal Component Analysis (PCA)	89%
6	Proposed Method	Schizophrenia Spectrum Disorders	Word2Vec + RF + Hyperparameter Tuning	93,81 % (SkipGram)

4. Conclusions

This research aims to detect speech patterns as one of the main symptoms in people with Schizophrenia (SZ) using the Word2Vec method, with a focus on comparing Continuous Bag of Words (CBOW) and Skip-Gram (SG) architectures. The results showed that the third experiment, which used hyperparameter tuning (vector size = 300, window = 5, min count = 2), gave the most outstanding results. In this experiment, the SG architecture achieved the highest accuracy of accuracy of 93% based on simple metrics, 98.21% based on ROC Curve, and 94% based on Cross-Validation, demonstrating its superior ability to capture complex semantic patterns in the SZ speech dataset compared to CBOW. The success of SG in this experiment is supported by its advantage of learning word context in depth, allowing the model to more accurately represent the semantic relations typical of SZ sufferers. This study confirms that parameter optimization not only improves model performance but also enables Word2Vec to be more effective in analyzing language patterns in the clinical domain. These findings make an important contribution to the development of Natural Language Processing (NLP) in clinical applications, such as early detection and monitoring of patients with schizophrenia. By utilizing this approach, it is expected that NLP-based systems can become a reliable support tool for healthcare professionals in the diagnosis and management of mental disorders. Further research can integrate other NLP models and expand the dataset to increase the generalizability and usefulness of this technology in clinical practice.

References

[1] Badan Pusat Statistik, *Profil Statistik Kesehatan*, vol. 7. 2023. Accessed: Aug. 14, 2024. [Online]. Available: https://www.bps.go.id/id/publication/2023/12/20/feffe5519c812d560bb131ca/profil-statistik-kesehatan-2023.html

[2] American Psychiatric Association, *Diagnostic and Statistical Manual of DSM-5TM*. 2013. Accessed: Aug. 14, 2024. [Online]. Available: https://www.psychiatryonline.org/dsm

[3] D. I. Velligan and S. Rao, "The Epidemiology and Global Burden of Schizophrenia," 2023, *Physicians Postgraduate Press Inc.* doi: 10.4088/JCP.MS21078COM5.

[4] T. Onitsuka *et al.*, "PCN FRONTIER REVIEW PCN Toward recovery in schizophrenia: Current concepts, findings, and future research directions," *Psychiatry Clin Neurosci*, vol. 76, no. 7, 2022, doi: 10.1111/pcn.13342/full.

[5] M. Pauzi, "Hubungan Beban Sosial dengan Kemampuan Keluarga Merawat Pasien Skizofreenia Pasca Pasung di Wilayah Kabupaten Bungo-Jambi," *Jurnal Inovasi Penelitian*, vol. Vol.2 No.5, 2021, Accessed: Jan. 25, 2025. [Online]. Available: https://ejournal.stpmataram.ac.id/JIP/article/view/915

[6] A. J. McGuinness *et al.*, "A systematic review of gut microbiota composition in observational studies of major depressive disorder, bipolar disorder and schizophrenia," Apr. 01, 2022, *Springer Nature*. doi: 10.1038/s41380-022-01456-3.

[7] J. A. Cortes-Briones, N. I. Tapia-Rivas, D. C. D'Souza, and P. A. Estevez, "Going deep into schizophrenia with artificial intelligence," *Schizophr Res*, vol. 245, pp. 122–140, Jul. 2022, doi: 10.1016/j.schres.2021.05.018.

[8] X. Chen, D. G. Chen, Z. Zhao, J. Zhan, C. Ji, and J. Chen, "Artificial image objects for classification of schizophrenia with GWAS-selected SNVs and convolutional neural network," *Patterns*, vol. 2, no. 8, Aug. 2021, doi: 10.1016/j.patter.2021.100303.

[9] X. Chen, H. Xie, and X. Tao, "Vision, status, and research topics of Natural Language Processing," *Natural Language Processing Journal*, vol. 1, p. 100001, 2022, doi: 10.1016/j.nlp.2022.100001.

[10] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimed Tools Appl*, vol. 82, no. 3, pp. 3713–3744, Jan. 2023, doi: 10.1007/s11042-022-13428-4.

[11] G. Di Gennaro, A. Buonanno, and F. A. N. Palmieri, "Considerations about learning Word2Vec," *Journal of Supercomputing*, vol. 77, no. 11, pp. 12320–12335, Nov. 2021, doi: 10.1007/s11227-021-03743-2.

[12] P. F. Muhammad, R. Kusumaningrum, and A. Wibowo, "Sentiment Analysis Using Word2vec and Long Short-Term Memory (LSTM) for Indonesian Hotel Reviews," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 728–735. doi: 10.1016/j.procs.2021.01.061.

[13] H. Xia, "Continuous-bag-of-words and Skip-gram for word vector training and text classification," in *Journal of Physics: Conference Series*, Institute of Physics, 2023. doi: 10.1088/1742-6596/2634/1/012052.

[14] H. Jayadianti, B. A. Arianti, N. H. Cahyana, S. Saifullah, and R. Dreżewski, "Improving sentiment analysis on PeduliLindungi comments: a comparative study with CNN-Word2Vec and integrated negation handling," *Science in Information Technology Letters*, vol. 4, no. 2, pp. 75–89, Nov. 2023, doi: 10.31763/sitech.v4i2.1184.

[15] S. Al-Saqqa, A. Awajan, and B. Hammo, "Performance Comparison of Word2Vec Models for Detecting Arabic Hate Speech on Social Networks," in *2022 International Conference on Emerging Trends in Computing and Engineering Applications (ETCEA)*, IEEE, Nov. 2022, pp. 1–5. doi: 10.1109/ETCEA57049.2022.10009734.

[16] S. C. Pereira, A. M. Mendonça, A. Campilho, P. Sousa, and C. Teixeira Lopes, "Automated image label extraction from radiology reports — A review," Mar. 01, 2024, *Elsevier B.V.* doi: 10.1016/j.artmed.2024.102814.

[17] A. E. Voppel, J. N. De Boer, S. G. Brederoo, H. G. Schnack, and I. E. C. Sommer, "Semantic and Acoustic Markers in Schizophrenia-Spectrum Disorders: A Combinatory Machine Learning Approach," *Schizophr Bull*, vol. 49, pp. S163–S171, Mar. 2023, doi: 10.1093/schbul/sbac142.

[18] F. Tsiwah, A. Mayya, and A. van Craneburgh, "Semantic-based NLP techniques discriminate schizophrenia and Wernicke's aphasia based on spontaneous speech Tsiwah," May 2024. Accessed: Dec. 27, 2024. [Online]. Available: https://research.rug.nl/en/publications/semantic-based-nlp-techniques-discriminate-schizophrenia-and-wern

[19] B. TaghiBeyglou and F. Rudzicz, "Context is not key: Detecting Alzheimer's disease with both classical and transformer-based neural language models," *Natural Language Processing Journal*, vol. 6, p. 100046, Mar. 2024, doi: 10.1016/j.nlp.2023.100046.

[20] Yuyun, A. D. Latief, T. Sampurno, Hazriani, A. O. Arisha, and Mushaf, "Next Sentence Prediction: The Impact of Preprocessing Techniques in Deep Learning," in *Proceedings - 2023 10th International Conference on Computer, Control, Informatics and its Applications: Exploring the Power of Data: Leveraging Information to Drive Digital Innovation, IC3INA 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 274–278. doi: 10.1109/IC3INA60834.2023.10285805.

[21] Rianto, A. B. Mutiara, E. P. Wibowo, and P. I. Santosa, "Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation," *J Big Data*, vol. 8, no. 1, Dec. 2021, doi: 10.1186/s40537-021-00413-1.

[22] J. Daniel and J. H. Martin, *Regular Expression, Text Normalization, Edit Distance*. 2023. Accessed: Aug. 16, 2024. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/old_jan23/2.pdf

[23] A. Chadha and B. Kaushik, "A Hybrid Deep Learning Model Using Grid Search and Cross-Validation for Effective Classification and Prediction of Suicidal Ideation from Social Network Data," *New Gener Comput*, vol. 40, no. 4, pp. 889–914, Dec. 2022, doi: 10.1007/s00354-022-00191-1.

[24] Q. Song *et al.*, "Optimizing Word Embeddings for Patient Portal Message Datasets with a Small Number of Samples," May 15, 2024. doi: 10.21203/rs.3.rs-4350387/v1.

[25] S. Jaradat, R. Nayak, A. Paz, and M. Elhenawy, "Ensemble Learning with Pre-Trained Transformers for Crash Severity Classification: A Deep NLP Approach," *Algorithms*, vol. 17, no. 7, Jul. 2024, doi: 10.3390/a17070284.

[26] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Jan. 2013, [Online]. Available: http://arxiv.org/abs/1301.3781

[27] L. Breiman, "Random Forests," Netherlanda, 2001. Accessed: Aug. 16, 2024. [Online]. Available: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

[28] T. Yu, W.-Z. Pei, C.-Y. Xu, C.-C. Deng, and X.-L. Zhang, "Identification of male schizophrenia patients using brain morphology based on machine learning algorithms," *World J*

Psychiatry, vol. 14, no. 6, pp. 804–811, Jun. 2024, doi: 10.5498/wjp.v14.i6.804.

[29] V. R. Gashkarimov, R. I. Sultanova, I. S. Efremov, and A. Asadullin, "Machine Learning Techniques in Diagnostics and Prediction of the Clinical Features of Schizophrenia: A Narrative Review," 2023, *Eco-Vector LLC*. doi: 10.17816/CP11030.

[30] Y. T. Jo, S. W. Joo, S. H. Shon, H. Kim, Y. Kim, and J. Lee, "Diagnosing schizophrenia with network analysis and a machine learning method," *Int J Methods Psychiatr Res*, vol. 29, no. 1, Mar. 2020, doi: 10.1002/mpr.1818.

[31] C. Cortes, V. Vapnik, and L. Saitta, "Support-Vector Networks Editor," Kluwer Academic Publishers, 1995. doi: 10.1007/BF00994018.

[32] B. Firmanto, H. Soekotjo, and H. Suyono, "Perbandingan Kinerja Algoritma Promethee dan Topsis Untuk Pemilihan Guru Teladan," *http://jurnal.unram.ac.id/index.php/jpp-ipa*, 2016, [Online]. Available: http://jurnal.unram.ac.id/index.php/jpp-ipa

[33] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit*, vol. 30, no. 7, pp. 1145–1159, Jul. 1997, doi: 10.1016/S0031-3203(96)00142-2.

[34] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," 1995. [Online]. Available: http//roboticsStanfordedu/"ronnyk

[35] D. Harris-Birtill and R. Harris-Birtill, "Understanding computation time," 2021.

[36] A. Sabina Uban, A. Maria Cristea, A. Dinu, L. P. Dinu, S. Georgescu, and iu Zoicas, "CoToHiLi at LSCDiscovery: the Role of Linguistic Features in Predicting Semantic Change," 2022. [Online]. Available: https://github.com/artetxem/vecmap

[37] T. P. Adewumi, F. Liwicki, and M. Liwicki, "Word2Vec: Optimal Hyper-Parameters and Their Impact on NLP Downstream Tasks," Mar. 2020, [Online]. Available: http://arxiv.org/abs/2003.11645