VOL. 11. NO. 2 NOVEMBER 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v11i2.6051

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

WORD2VEC OPTIMALIZATION USING TRANSFER LEARNING IN INDONESIAN LANGUAGE FOR HIGHER EDUCATION

Dwiza Riana[1]; Sri Hadianti[1*]; Herdian Tohir[1]; Jarwadi[2]; Tjaturningsih Rosdiana[2]; Evi Sopandi[2]; Dinar Ajeng Kristiyanti[3]

Departemen of Technology Information[1]
Universitas Nusa Mandiri[1]
https://www.nusamandiri.ac.id/[1]
dwiza@nusamandiri.ac.id, sri.shv@nusamandiri.ac.id*, htohir.ht@gmail.com

BRIN Education Research Center[2]
Badan Riset dan Inovasi Nasional[2]
https://ipsh.brin.go.id/[2]
jarwadi10@gmail.com, trosdiana.27@gmail.com, evi.sopandi@brin.go.id

Department of Information System[3]
Universitas Multimedia Nusantara[3]
https://www.umn.ac.id/[3]
dinar.kristiyanti@umn.ac.id

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *Natural language processing (NLP) in Indonesian faces challenges due to limited linguistic resources, particularly in developing optimal word embedding models. This study optimizes the Word2Vec model for Indonesian in higher education contexts by leveraging transfer learning and lexicon expansion. Using a dataset of 4,463 higher education related tweets consisting of positive and negative sentiment categories, the proposed NewWord2Vec model combined with a Support Vector Machine (SVM) classifier achieved a 4% improvement in word detection accuracy compared to the standard Word2Vec. This enhancement demonstrates better performance in capturing linguistic nuances and sentiment orientation in Indonesian text. However, the model's applicability remains limited to higher education terminology, and potential biases from transfer learning must be addressed. Future research should expand the dataset to diverse domains and refine the transfer learning process to better capture contextual variations in Indonesian. These findings contribute to advancing NLP applications in Indonesian, particularly for automated assessment systems, recommendation tools, and academic decision-making processes.*

Keywords: *indonesian language, NLP, Word2Vec, transfer learning, optimization*

Intisari— *Pemrosesan bahasa alami (NLP) dalam bahasa Indonesia menghadapi tantangan karena keterbatasan sumber daya linguistik, khususnya dalam pengembangan model penyisipan kata (word embedding) yang optimal. Penelitian ini mengoptimalkan model Word2Vec untuk konteks pendidikan tinggi dengan memanfaatkan transfer learning dan perluasan leksikon. Dengan menggunakan dataset berjumlah 4.463 tweet terkait pendidikan tinggi yang terdiri dari kategori sentimen positif dan negatif, model NewWord2Vec yang dikombinasikan dengan pengklasifikasi Support Vector Machine (SVM) menunjukkan peningkatan akurasi deteksi kata sebesar 4% dibandingkan dengan Word2Vec standar. Peningkatan ini menunjukkan kinerja yang lebih baik dalam menangkap nuansa linguistik dan orientasi sentimen pada teks berbahasa Indonesia. Namun, penerapan model ini masih terbatas pada terminologi pendidikan tinggi, dan potensi bias dari proses transfer learning perlu diatasi. Penelitian selanjutnya disarankan untuk memperluas*

VOL. 11. NO. 2 NOVEMBER 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i2.6051

dataset ke berbagai domain dan menyempurnakan proses transfer learning agar lebih mampu menangkap variasi kontekstual dalam bahasa Indonesia. Temuan ini berkontribusi dalam memajukan aplikasi NLP berbahasa Indonesia, khususnya untuk sistem penilaian otomatis, alat rekomendasi, dan proses pengambilan keputusan akademik.

***Kata Kunci**: Bahasa Indonesia, NLP, Word2Vec, transfer learning, optimasi*

INTRODUCTION

In recent years, there has been a growing interest in understanding and analyzing public opinions on higher education, particularly through social media platforms [1]. This trend reflects increasing concerns about the quality, accessibility, and relevance of higher education institutions in contemporary society [2]. Surveys indicate that a significant portion of individuals actively participate in discussions, sharing their experiences, perceptions, and critiques of higher education on social media [3]. Facilitated by information technology, smartphone users express their views in their native languages, each with distinct linguistic structures and contextual nuances [4]. These opinions vary based on individuals' background knowledge, experiences, and values, making them valuable sources for understanding educational contexts.

Natural Language Processing (NLP) has played a crucial role in analyzing textual data from social media, enabling automated sentiment analysis, opinion mining, and topic modeling [5]. NLP techniques, including machine learning and deep learning models, have been widely applied in various domains such as healthcare, finance, and customer feedback analysis [6]. However, research in NLP for education, particularly in analyzing sentiment and behavioral patterns related to learning success in higher education, remains limited. This gap highlights the need for further exploration of sentiment analysis in educational contexts, particularly in understanding how students and other stakeholders perceive learning success and institutional effectiveness.

Social media has been extensively utilized for sentiment analysis across various domains, including studies on COVID-19, politics, tourism, products, and applications. Sentiment-based behavioral analysis supports decision-making processes in these areas and continues to evolve. Within education, studies have used opinion mining to examine student responses in elementary and secondary schools, improve teacher evaluations, and automate student feedback analysis [7], Despite these advancements, most studies rely on standard Word2Vec datasets, which often lack the vocabulary richness necessary for domain-specific applications in higher education.

The Word2Vec model has been widely adopted in NLP for generating word embeddings, but it has notable limitations, including a constrained vocabulary that affects its ability to process diverse linguistic expressions in domain-specific contexts [8]. Research employing Word2Vec often faces challenges in identifying and representing words accurately, particularly in resource-limited languages such as Indonesian. The Support Vector Machine (SVM) method has demonstrated effectiveness in handling complex data, while deep learning models provide more precise and interpretable word representations [9], Additionally, transfer learning has emerged as a solution to address vocabulary limitations by leveraging knowledge from resource-rich languages, enhancing word recognition and semantic understanding [10].

Several studies have explored the integration of transfer learning in NLP, demonstrating its effectiveness in improving word representation and model accuracy across different domains [11], [12]. However, research applying transfer learning to optimize Word2Vec for sentiment analysis in Indonesian higher education discourse remains scarce.

The proposed method outlined in this paper furnishes a systematic approach to optimizing the Word2vec model for the Indonesian language while tackling the unique challenges posed by the higher education domain. This research constitutes a significant contribution by furnishing a refined methodology tailored to the Indonesian context, thereby augmenting the accuracy and relevance of sentiment analysis and opinion mining in higher education discourse on social media platforms.

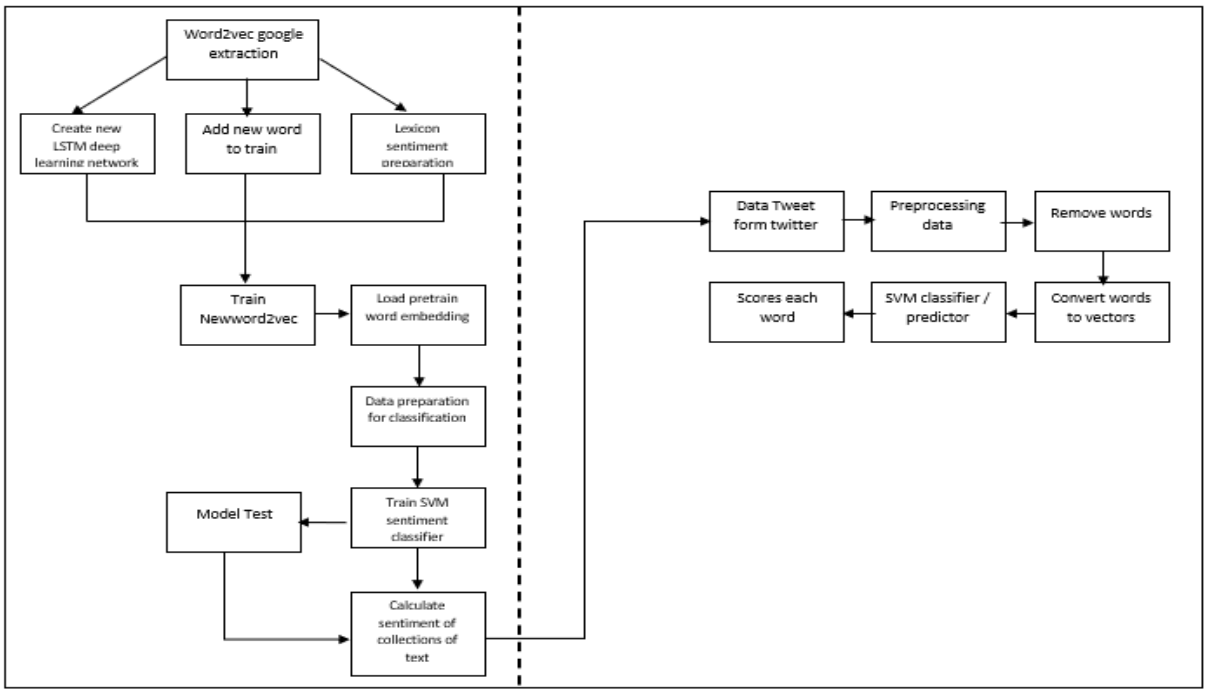
MATERIALS AND METHODS

The research conducted in this study includes several stages. First, extract the Word2Vec model from Google, then create a new LSTM deep learning network. This is followed by adding new words for training and preparing a sentiment lexicon. With the obtained data, the NewWord2Vec model is trained. The next dataset is loaded using pre-trained word embeddings. The following step is

data preparation for classification. Then, the SVM sentiment classifier is trained, its results are tested on the model, and the sentiment of text collections is calculated. This process constitutes the stages of network optimization. The resulting data from network optimization is implemented into Twitter data by crawling tweet data. The research design can be seen in Figure 1.

A. Google Word2Vec Extraction

This phase of the research involves a series of methodical operations beginning with the acquisition of the pre-trained Word2Vec model from Google's repository. Following this the model is loaded into memory, facilitating efficient access and manipulation. Once the model is active, vector representations of words related to the concept of success in higher education are retrieved. These word vectors are subjected to various operations including vector arithmetic and similarity assessments, to explore the complex relationships between terms [13]. Subsequently, a thorough evaluation and adjustment phase is conducted to enhance the model's performance and ensure its relevance to the higher education success context. This meticulous process yields a refined set of words and their corresponding vector representations, which serve as foundational material for subsequent research and analysis.



Source: (Research Results, 2025)

Figure 1. Research Design

B. New LSTM Deep Learning Network

This procedure encompasses the development of a sophisticated deep learning Long Short-Term Memory (LSTM) model, characterized by careful selection of the appropriate platform and libraries to construct the neural network architecture. Following the selection phase, the requisite libraries are imported into the development environment, and comprehensive datasets are gathered to facilitate both training and evaluation phases of the LSTM model. Subsequently, meticulous configuration of the LSTM model is performed, involving fine-tuning parameters such as network depth, hidden layer dimensions, and learning rates [14]. Finally, the LSTM model undergoes rigorous evaluation, where its performance metrics are scrutinized to ascertain its efficacy in capturing temporal dependencies and patterns within the dataset, thus ensuring its suitability for real-world applications in sequential data analysis tasks.

C. Sentiment Lexicon Preparation

In this crucial stage of sentiment analysis, a meticulous selection and refinement of dictionaries are undertaken, involving the extraction of negative and positive polarities from extensive lexicon data. The process begins with the careful curation of lexicons to ensure that the dictionaries are comprehensive and contextually relevant. Words are then systematically grouped into distinct sentiment categories, such as positive, negative, and

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 2 NOVEMBER 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i2.6051

neutral, with special attention given to handling compound words, idiomatic expressions, and phrases that may carry specific contextual meanings [15]. This includes considering the subtleties and nuances of words that might have conflicting sentiments based on their usage in different contexts.

Advanced techniques are employed to manage polysemy and semantic ambiguities, ensuring that words with multiple meanings are accurately categorized according to their sentiment in given contexts. Additionally, the integration of context-aware algorithms helps adjust sentiment scores based on surrounding words, thereby enhancing the precision of sentiment detection.

Special attention is also given to cultural and linguistic variations, ensuring that the sentiment analysis model is robust and adaptable to diverse linguistic contexts. This comprehensive and methodical approach guarantees that the sentiment analysis is not only accurate but also capable of capturing the complexity of human emotions and expressions across different domains and languages. The end result is a highly refined and reliable sentiment analysis framework that provides valuable insights into the emotional tone of textual data.

D. New Words for Training

In this phase, all Indonesian words are input into the system and represented using the Two Hot Encoding (THE) technique, a modified version of the conventional One-Hot Encoding [16]. Unlike One-Hot Encoding, which activates only one binary position for each word, THE activates two positions simultaneously one representing the target word and another representing its semantically closest word determined from the pre-trained Word2Vec embedding space. This dual activation allows the model to retain both the identity of the word and partial information about its contextual similarity, thereby integrating a lightweight semantic relationship into the encoding process. The encoded vectors generated through THE are then used to enrich the training data for the NewWord2Vec model.

Table 1. One Example Of The Results From THE

Word	Two Hot Encoding (THE)	Target Vectors
Memenangkan (Winner)	0 0 0 0 0 0	0.0488 -0.0012 -0.0405 0.1073
	0 0 0 0 0 0	-0.0110 0.0336 -0.0138 0.0059 0.0149 0.0486 0.0898 0.0738
	0 0 0 0 1 1	0.3153 -0.1003 0.0276 0.0139 0.1194 0.0982

Source: (Research Results, 2025)

E. Training NewWord2Vec

NewWord2Vec extends the capabilities of Word2Vec by incorporating Indonesian vocabulary. During this stage, the network undergoes training with a configuration tailored for the words generated in the third phase. This entails adjusting parameters such as vector size, learning algorithms, and preprocessing techniques optimized specifically to obtain richer word vector representations suitable for the Indonesian language [17]. The training process may also involve additional adjustments to accommodate the complexity and variations in the structure of the Indonesian language.

F. Loading Pre-trained Word Embeddings

In this phase, the process entails integrating pre-existing words that have undergone training. The objective is to leverage the vectorized representations of these words acquired through natural language processing (NLP) techniques [18]. By incorporating these learned word vectors into the analysis, the aim is to enhance outcomes and mitigate the risk of overfitting, thus refining the model's performance and ensuring its adaptability across various linguistic contexts. This integration enables the model to capture nuanced semantic relationships and nuances present within the dataset, thereby fostering more accurate and nuanced analyses in natural language processing tasks.

G. Data Preparation for Classification

The data is segmented into training and testing datasets, with 90% allocated for training and 10% for testing purposes, specifically tailored for the SVM model used in sentiment analysis [19]. These datasets are securely stored within a designated repository, ensuring easy access and proper management throughout the analysis pipeline. This segregation ensures that the model is trained on a substantial portion of the data while retaining a separate subset for evaluating its performance, ultimately enhancing the robustness and reliability of the sentiment analysis system.

H. Calculating Sentiment of Collections of Text

Sentiment analysis is conducted on Twitter tweet data individually for each tweet, employing a combination of network models and Support Vector

Machines (SVM) [20]. The dataset utilized originates from the Google Word2Vec extraction lexicon database for Indonesian, with modifications made to accommodate new vocabulary arrangements, including additions and subtractions. This comprehensive approach ensures a thorough examination of sentiment across diverse linguistic contexts, facilitating a nuanced understanding of public opinion and discourse on the Twitter platform.

I. Twitter Tweet Data

During this phase, an automated process was implemented to crawl Twitter for relevant discussions about success in higher education [21]. The data collection covered the period from January to December 2023, capturing both real-time and historical tweets. A total of 11 Indonesian keywords including tuition fees, thesis guidance, and online lectures were used to query Twitter's API. To ensure linguistic consistency, language filtering was applied using the langdetect Python library to retain only tweets written in Indonesian, while code-mixed or foreign-language tweets were excluded. The cleaning process involved several steps: removing URLs, emojis, user mentions (@), hashtags, punctuation, and duplicate entries. Additionally, text normalization was performed by converting all text to lowercase and eliminating common stopwords using the NLTK Indonesian stopword list. The resulting 4,463 cleaned tweets were compiled into a structured database for subsequent preprocessing and sentiment analysis.

J. Preprocessing Data

In this phase, the data obtained from crawling Twitter undergoes a series of preprocessing steps aimed at refining its quality and preparing it for subsequent analysis. The process commences with tokenization, which involves segmenting the text into 16 distinct tokens, facilitating granular analysis of the content [21]. Subsequently, punctuation marks are systematically eliminated from the text, as they often do not substantially augment the semantic meaning of the discourse in many natural language processing scenarios. Furthermore, stop words, though absent in this specific tweet, are typically removed to eliminate common linguistic artifacts that may hinder analysis [22]. Additionally, all text is converted to lowercase to standardize the data format, ensuring consistency across the dataset [23]. These preprocessing steps collectively enhance the readability and interpretability of the data, streamlining subsequent analytical procedures and facilitating more accurate insights extraction.

K. Removing Words

During this stage, the focus shifts to filtering out words that are absent in the Word2Vec or NewWord2Vec databases, ensuring that only relevant and meaningful tokens remain for further analysis. As a consequence, the initial set of 13 tokens is significantly pruned down to 5 tokens, as 8 tokens were not found within the Word2Vec or NewWord2Vec database. This proactive curation process is distinct from the preceding phase, which primarily targeted the removal of stop words to streamline the data [24]. By eliminating words that lack corresponding embeddings in the database, this approach enhances the quality and relevance of the dataset, thereby fostering more accurate and insightful analyses in subsequent stages of the research.

L. Converting Words to Vectors

In this stage, the transformation of words into numerical representations is undertaken to facilitate computer comprehension and manipulation of textual data in a structured numerical format. This conversion process enables computers to efficiently process and analyze textual information, laying the groundwork for various natural language processing (NLP) tasks [25]. Furthermore, this stage encompasses the assessment of similarity and relationships between words derived from the preprocessed data. By quantifying the semantic relationships and similarities between words, insights into the underlying structures and associations within the dataset are unveiled, empowering more nuanced and sophisticated analyses. Through these combined efforts, the stage bridges the gap between linguistic concepts and computational frameworks, enabling a deeper understanding of textual data and facilitating advanced NLP applications.

M. SVM Classifier/Predictor

The SVM classifier/predictor is a type of machine learning algorithm used for classification and regression tasks. In classification tasks, such as sentiment analysis, SVM categorizes data points into different classes based on their features. In the initial training stage, the SVM algorithm is trained on a labeled dataset. Each data point in the training dataset is represented by a set of features, and each data point is assigned to a specific class (positive or negative sentiment) [26]. During the training phase, SVM seeks to find the optimal hyperplane that separates the data points belonging to different

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 2 NOVEMBER 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i2.6051

classes. After the SVM model is trained, it can be used to predict the class labels of new, unseen data points. The SVM algorithm evaluates the input features of new data points and determines on which side of the hyperplane they lie, assigning the data points to one of the classes accordingly. I chose the SVM method because it has a strong ability to handle high-dimensional data and find the optimal hyperplane that separates classes with maximum margin, resulting in good generalization on unseen data. Additionally, SVM uses the kernel trick to handle non-linearly separable data, making it effective for a variety of classification and regression tasks, such as sentiment analysis. This method is also more resistant to overfitting, especially on small or noisy datasets, and has been proven to perform well in various real-world applications.

N. Scoring Each Word

The output of the sentiment analysis provides a positive/negative score from the average of the previous process. For example, if the value (score) is greater than 0, the sentiment is considered positive; if the score is less than 0, the sentiment is negative, and 0 represents neutral sentiment [27]. From the two examples above for Word2Vec and NewWord2Vec, we can conclude that the differences lie in how they convert words to vectors, which depends on their vocabulary databases. The final stage is to calculate the average score from all groups of text in a tweet with unique keywords.

RESULTS AND DISCUSSION

The experimental analysis contrasting Word2Vec with NewWord2Vec elucidated a discernible discrepancy, wherein NewWord2Vec exhibited a notable capacity to detect approximately 4% more lexical items compared to its predecessor, the original Word2Vec model. This substantial augmentation in word detection proficiency was meticulously attained through the rigorous training of NewWord2Vec utilizing an expanded lexicon specifically curated to align with the nuanced intricacies of the higher education milieu. Notably, the training corpus for NewWord2Vec encompassed a meticulously curated dataset comprising 2,926 distinct entries, categorized into 1,100 instances denoting positive sentiment and 1,826 instances indicative of negative sentiment. The tabulated data in Table 2 delineates the quantitative disparities in word detection capabilities manifested by both models across a spectrum of keywords relevant to higher education.

The sentiment analysis scores for the tweets are meticulously archived within an Excel spreadsheet, where each score is calibrated within the range of -1 to 1. Within this scale, positive scores indicate affirmative sentiments, whereas negative scores denote pessimistic sentiments, and a score of zero signifies neutral sentiment. These scores are computed on a per-tweet basis, with the average sentiment score calculated for each keyword, enabling a comprehensive assessment of prevailing sentiments. The process entails meticulous computation of sentiment scores for individual words, which are subsequently aggregated to derive the sentiment score for each tweet, culminating in the summation of all sentiment scores corresponding to tweets pertinent to a specific keyword. This process is succinctly encapsulated within the tabular representation in Table 3, further augmented by the elucidatory visualization in Figure 2.

The sentiment scores for each tweet were computed and synthesized, as illustrated in Table 4. Each score represents the sentiment analysis outcome of a tweet, with positive values indicating positive sentiment and negative values indicating negative sentiment. The aggregated average score for each keyword was utilized to draw conclusions regarding the overall sentiment within the dataset. This methodological approach facilitated the extraction of nuanced insights into sentiment dynamics, contributing to a comprehensive understanding of public sentiment within the context of the analyzed Twitter data.

Table 2. Detected Words For Each Network

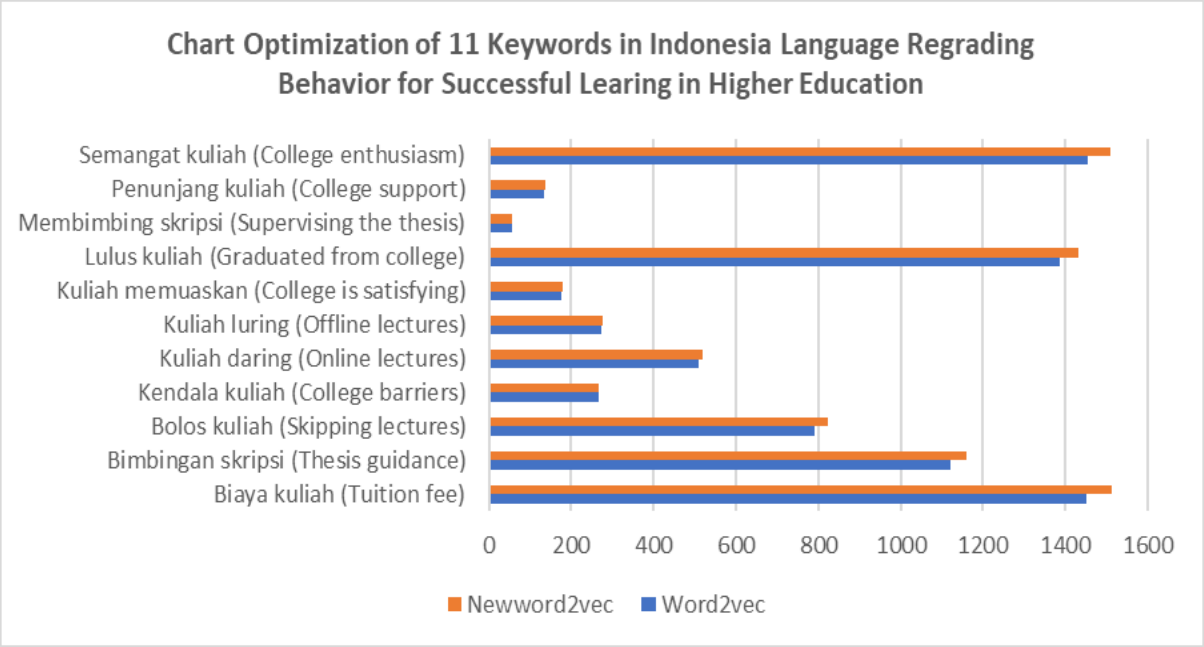
Keywords	Word2vec	Newword2vec
Biaya kuliah (Tuition fee)	1,452	1,512
Bimbingan skripsi (Thesis guidance)	1,122	1,160
Bolos kuliah (Skipping lectures)	792	823
Kendala kuliah (College barriers)	266	268
Kuliah daring (Online lectures)	511	518
Kuliah luring (Offline lectures)	272	276
Kuliah memuaskan (College is satisfying)	175	180
Lulus kuliah (Graduated from college)	1,388	1,432
Membimbing skripsi (Supervising the thesis)	55	57
Penunjang kuliah (College support)	133	138
Semangat kuliah (College enthusiasm)	1,455	1,511

Source: (Research Results, 2025)

Table 3. Optimization Indonesian Language

Keywords	Optimization (%)
Biaya kuliah (Tuition fee)	4
Bimbingan skripsi (Thesis guidance)	3
Bolos kuliah (Skipping lectures)	4
Kendala kuliah (College barriers)	1
Kuliah daring (Online lectures)	1
Kuliah luring (Offline lectures)	1
Kuliah memuaskan (College is satisfying)	3
Lulus kuliah (Graduated from college)	3
Membimbing skripsi (Supervising the thesis)	4
Penunjang kuliah (College support)	4
Semangat kuliah (College enthusiasm)	4

Source: (Research Results, 2025)



Source: (Research Results, 2025)

Figure 2. Visualization Optimization

Table 4. Optimization Indonesian Language

Score	Tweets
0.675352725	Satoru dan Nobara as siblings beda 8 tahun. Satoru kerja di luar kota dan tinggal di kontrakan. Waktu Nobara mo kuliah https://t.co/RjLHpXCkdu
0.245822891	RT @GoDisko: Soal koperasi simpan pinjam, jadi inget cerita OB yang putus kuliah karena uangnya dia pertaruhkan di usaha koperasi simpan
0.474983051	Soal koperasi simpan pinjam, jadi inget cerita OB yang putus kuliah karena uangnya dia pertaruhkan di usaha koperasi https://t.co/JyAfdPtU9c
0.283322547	gila pontang panting aku nyari duit buat biaya hidup sm kuliah malah ga dicairin sama sekali duitku, shopee na-jisss https://t.co/uwoqN7sGl0
-0.026310111	@tanyakanrl nder beneran minta ganti? mungkin mksd beliau spya u kuliah pinter", g neko", lulus tepat waktu dn biar https://t.co/Iut7Byhxo6
0.592488515	RT @tanyakanrl: Tanyarl emang mahar harus sesuai / lebih sama biaya kuliah ya? https://t.co/ZqhqQCGHXP
0.894895116	@drkoko28 Sebelum menentukan jurusan kuliah, coba riset biaya kuliahnya
0.816794823	@faetdino @tanyakanrlOhiyasatulg. Sekolah tu tanda kita prnh belajar
0.875708091	@kepikhijauu @tanyakanrl Kalo gk salah dulu pernah liat biaya kuliah SM Unpad kedokteran sampe 50jt *CMIIW gk tau https://t.co/IHb5VKuv7M

Source: (Research Results, 2025)

This study evidences that the integration of NewWord2Vec enhances word detection and sentiment analysis accuracy through the incorporation of a broader Indonesian lexicon, meticulously tailored to the realm of higher education. While NewWord2Vec demonstrates superior performance over Word2Vec, future research should explore its scalability across larger datasets and varying domains. Additionally, further refinements in training methodologies and lexicon expansion may yield greater improvements in sentiment classification accuracy. These findings underscore the importance of continuously refining natural language processing techniques to better capture linguistic nuances in domain-specific contexts.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 2 NOVEMBER 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i2.6051

CONCLUSION

This study presents an optimization of the Word2Vec model for the Indonesian language in higher education contexts through the integration of transfer learning and lexicon expansion. The NewWord2Vec model, by improving word detection and contextual representation, contributes to more precise sentiment classification and a richer understanding of public discourse in the education domain. The enhanced lexical coverage allows the model to capture subtle linguistic cues that standard Word2Vec often overlooks, resulting in more reliable identification of emotional tone and semantic intent. Beyond numerical gains, this advancement demonstrates the importance of domain-specific embedding models for under-resourced languages like Indonesian. Nevertheless, the model's focus on higher education terminology limits its generalization to broader topics. Future research should expand training datasets across multiple domains and explore integration with contextual models such as BERT or RoBERTa to achieve deeper semantic comprehension and greater cross-domain adaptability. These efforts can strengthen NLP applications in automated feedback systems, educational analytics, and intelligent decision-support tools..

ACKNOWLEDGMENT

The author would like to thank the Educational Transformation Fund to Produce Future Lead- ers Collaborative Research Initiative (DRIVEN) 2022 implemented by BrinNational Research and Innovation Agency, Tanoto Foundation and The Conversation for supporting this research through the "Word2vec Optimalization in Indonesian Language for Higher Education with Trans- fer Learning", 2022.

REFERENCE

[1] E. Shamsi and H. Bozorgian, "A review of empirical studies of social media on language learners' willingness to communicate," *Educ. Inf. Technol.*, vol. 27, no. 4, pp. 4473–4499, 2022, doi: 10.1007/s10639-021-10792-w.

[2] J. Fan, "Innovation and Exploration of the Path to Cultivating University Students' Cultural Awareness through New Media in the Context of Cultural Inheritance," *Media Commun. Res.*, vol. 4, no. 11, pp. 61–70, 2023, doi: 10.23977/mediacr.2023.041109.

[3] Y. Purnama and A. Asdlori, "The Role of Social Media in Students' Social Perception and Interaction: Implications for Learning and Education," *Technol. Soc. Perspect.*, vol. 1, no. 2, pp. 45–55, 2023, doi: 10.61100/tacit.v1i2.50.

[4] S. Kwayu, M. Abubakre, and B. Lal, "The influence of informal social media practices on knowledge sharing and work processes within organizations," *Int. J. Inf. Manage.*, vol. 58, no. December 2020, p. 102280, 2021, doi: 10.1016/j.ijinfomgt.2020.102280.

[5] A. Sandu, L. A. Cotfas, A. Stănescu, and C. Delcea, *A Bibliometric Analysis of Text Mining: Exploring the Use of Natural Language Processing in Social Media Research*, vol. 14, no. 8. 2024.

[6] N. A. Sharma, A. B. M. S. Ali, and M. A. Kabir, "A review of sentiment analysis: tasks, applications, and deep learning techniques," *Int. J. Data Sci. Anal.*, no. September, 2024, doi: 10.1007/s41060-024-00594-x.

[7] A. Namoun and A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," *Appl. Sci.*, vol. 11, no. 1, pp. 1–28, 2021, doi: 10.3390/app11010237.

[8] D. S. Asudani, N. K. Nagwani, and P. Singh, *Impact of word embedding models on text analytics in deep learning environment: a review*, vol. 56, no. 9. Springer Netherlands, 2023.

[9] U. Naseem, I. Razzak, S. K. Khan, and M. Prasad, "A Comprehensive Survey on Word Representation Models: From Classical to State-of-the-Art Word Representation Language Models," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 20, no. 5, pp. 1–46, 2021, doi: 10.1145/3434237.

[10] M. Krichen, "Convolutional Neural Networks: A Survey," *Computers*, vol. 12, no. 8, pp. 1–41, 2023, doi: 10.3390/computers12080151.

[11] T. Adimulam, S. Chinta, and S. K. Pattanayak, "Transfer Learning in Natural Language Processing : Overcoming Low-Resource Challenges," vol. 11, no. 2, pp. 65–79, 2022.

[12] M. Iman, H. R. Arabnia, and K. Rasheed, "A Review of Deep Transfer Learning and Recent Advancements," *Technologies*, vol. 11, no. 2, pp. 1–14, 2023, doi: 10.3390/technologies11020040.

[13] N. Nedjah, I. Santos, and L. de Macedo Mourelle, "Sentiment analysis using convolutional neural network via word embeddings," *Evol. Intell.*, vol. 15, no. 4, pp. 2295–2319, 2022, doi: 10.1007/s12065-

019-00227-4.

[14] T. T. J. Kiran and P. P. Jadhav, "Optimizing Deep Learning: Unveiling the Collective Wisdom of Swarm Intelligence for LSTM Parameter Tuning," *Int. J. Intell. Syst. Appl. Eng.*, vol. 12, no. 13s, pp. 432–439, 2024.

[15] M. A. Jassim, D. H. Abd, and M. N. Omri, "A survey of sentiment analysis from film critics based on machine learning, lexicon and hybridization," *Neural Comput. Appl.*, vol. 35, no. 13, pp. 9437–9461, 2023, doi: 10.1007/s00521-023-08359-6.

[16] P. Tschisgale, P. Wulff, and M. Kubsch, "Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory," *Phys. Rev. Phys. Educ. Res.*, vol. 19, no. 2, p. 20123, 2023, doi: 10.1103/PhysRevPhysEducRes.19.020123.

[17] A. Setyanto *et al.*, "Arabic Language Opinion Mining Based on Long Short-Term Memory (LSTM)," *Appl. Sci.*, vol. 12, no. 9, pp. 1–18, 2022, doi: 10.3390/app12094140.

[18] F. R. Zagatti, G. Y. Shimizu, and H. D. E. M. Caseli, "Investigating the Relationship Between Text Vectorization Cosine Similarity and Classification Performance," vol. 13, no. June, 2025, doi: 10.1109/ACCESS.2025.3595423.

[19] R. K. Halder *et al.*, "ML-CKDP: Machine learning-based chronic kidney disease prediction with smart web application," *J. Pathol. Inform.*, vol. 15, no. December 2023, p. 100371, 2024, doi: 10.1016/j.jpi.2024.100371.

[20] M. W.Habib and Z. N. Sultani, "Twitter Sentiment Analysis Using Different Machine Learning and Feature Extraction Techniques," *Al-Nahrain J. Sci.*, vol. 24, no. 3, pp. 50–54, 2021, doi: 10.22401/anjs.24.3.08.

[21] S. Nazir, M. Asif, M. Rehman, and S. Ahmad, "Machine learning based framework for fine-grained word segmentation and enhanced text normalization for low resourced language," no. 1, pp. 1–19, 2024, doi: 10.7717/peerj-cs.1704.

[22] J. Fehle, T. Schmidt, and C. Wolff, "Lexicon-based sentiment analysis in german: Systematic evaluation of resources and preprocessing techniques," *KONVENS 2021 - Proc. 17th Conf. Nat. Lang. Process.*, pp. 86–103, 2021.

[23] X. Tannier *et al.*, "Development and Validation of a Natural Language Processing Algorithm to Pseudonymize Documents in the Context of a Clinical Data Warehouse," *Methods Inf. Med.*, 2023, doi: 10.1055/s-0044-1778693.

[24] I. Biri, U. T. Kucuktas, F. Uysal, and F. Hardalac, "Forecasting the future popularity of the anti-vax narrative on Twitter with machine learning," *J. Supercomput.*, vol. 80, no. 3, 2024, doi: https://doi.org/10.1007/s11227-023-05567-8.

[25] M. A. Al-Garadi, Y. C. Yang, and A. Sarker, "The Role of Natural Language Processing during the COVID-19 Pandemic: Health Applications, Opportunities, and Challenges," *Healthc.*, vol. 10, no. 11, pp. 1–19, 2022, doi: 10.3390/healthcare10112270.

[26] L. A. Demidova, "Two-stage hybrid data classifiers based on svm and knn algorithms," *Symmetry (Basel).*, vol. 13, no. 4, 2021, doi: 10.3390/sym13040615.

[27] A. Mahmoudi, D. JEMIELNIAK, and L. CIECHANOWSKI, "Assessing Accuracy : A Study of Lexicon and Rule-Based Packages in R and Python for Sentiment Analysis," *IEEE Access*, vol. 12, no. February, pp. 20169–20180, 2024, doi: 10.1109/ACCESS.2024.3353692.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri