Khazanah Journal of Religion and Technology
Universitas Islam Negeri Sunan Gunung Djati Bandung, Indonesia
https://journal.uinsgd.ac.id/kjrt
Online ISSN: 2987-6060
Volume 3 Issue 2 (2025)
DOI: https://doi.org/10.15575/kjrt.v3i2.1748

Thematic Grouping of Quranic Verse Translations Based on Word2Vec and K-Means Clustering

Ahmad Badru Al Husaeni
Department of Informatics
UIN Sunan Gunung Djati Bandung
Bandung, Indonesia
badrualhusaeni23@gmail.com

Alif Firmansyah Putra
Bank BCA Syariah
Jakarta, Indonesia
aliffirmansh@gmail.com

Adi Purnama
Artristik Studio Bandung
Bandung, Indonesia
adipurnamaa8@gmail.com

Adly Juliarta Lerian
Department of Informatics
UIN Sunan Gunung Djati Bandung
Bandung, Indonesia
juliartalerian@gmail.com

Diman Fathurahman
Department of Informatics
UIN Sunan Gunung Djati Bandung
Bandung, Indonesia
dimanfn24@gmail.com

Abstract— This study aims to group thematically translated texts of Indonesian Quranic verses using a Word2Vec-based machine learning approach and the KMeans Clustering algorithm. The process begins with text preprocessing, creating vector representations using Word2Vec, and then clustering using KMeans with quality evaluation using the Silhouette Score metric. The experimental results show that the model is able to form six main thematic clusters that semantically describe themes such as prayer and hope, moral evil, social law, the teachings of revelation, divinity, and the stories of figures and ethics. Two-dimensional visualization with PCA strengthens the interpretation of the formed clustering patterns. This study proves that the unsupervised learning approach can be relied upon to support the automation of digital thematic interpretation objectively and systematically. In addition, the results of this clustering have the potential to become the basis for the development of topic-based verse search systems, contextual Quranic learning applications, and technology-based exploration of Islamic studies. This study also supports the achievement of Sustainable Development Goals (SDGs) point 4 regarding increasing access to inclusive and quality education through information technology.

Keywords- Al-Qur'an, Digital Interpretation, K-Means Clustering, Natural Language Processing, Thematic Clustering, Word2Vec

I. INTRODUCTION

The development of information technology in the current digital era has had a significant impact on various fields, including religious studies. One form of technology utilization in the Islamic field is to analyze religious texts using data-based computing methods. The Qur'an, as the holy book of Muslims, contains various moral teachings, laws, and social values conveyed in the form of verses. Each verse in the Qur'an has a diverse context, message, and meaning, which in practice is often grouped according to certain themes

to facilitate understanding. The process of grouping themes or topics of verses is known as thematic interpretation [1].

Until now, thematic studies of the Qur'an have generally been conducted manually by *mufassir* (interpreters) using a deductive method based on searching for specific keywords or topics. Although this method has proven effective, amidst the large number of verses reaching more than 6,000 in the Qur'an, this manual process requires a long time, expert personnel, and a high level of analytical consistency. This is where the role of machine learning technology can be optimized, especially unsupervised learning methods such as clustering, which allow the system to automatically group text data based on certain similarity patterns without the need for initial labels [2], [3], [4].

In the field of Natural Language Processing (NLP), one approach that is widely used to represent text into numeric form is word embedding. Word embedding is able to capture semantic relationships between words in a text corpus and present them in the form of fixed-dimensional vectors. One popular and widely used word embedding method is Word2Vec, developed by Google Research [5]. Word2Vec is able to map words with similar meanings into adjacent vector spaces. This model is divided into two main architectures, namely Continuous Bag of Words (CBOW) and Skip-Gram, each of which has a way of working to predict words from context or vice versa. In this study, the author uses Word2Vec to represent words in the Indonesian translation of the Quranic verses into vector form.

Furthermore, these vectors are used as input for the clustering process using the KMeans Clustering algorithm. KMeans is a simple yet effective centroid-based clustering algorithm for dividing data into groups based on the distance between vectors [6]. The best number of clusters is determined by measuring the Silhouette Score, an evaluation metric that measures the extent to which data matches its own cluster compared to other clusters [7], [8]. This study aims to identify whether translated Quranic verses can be automatically grouped based on the similarity of meaning or topic of their translation. The results of this clustering are expected to serve as an initial reference for the development of a digital thematic interpretation system, a topic-based verse search system, or even to support computational linguistics-based Quranic semantic analysis studies. In addition, this project also contributes to strengthening technology-based religious literacy and supports the goal of Sustainable Development Goals (SDGs) point 4, namely ensuring inclusive and equitable quality education for all, through more open and structured access to religious technology [9]. Previous research conducted by T. Mikolov et al. has proven that Word2Vec can effectively map semantic relationships between words in various languages [10]. The use of Word2Vec in the context of religious texts, particularly the Quran, is still relatively rarely explored, especially in Indonesian-language corpora.

Therefore, this study attempts to bridge this gap by testing the effectiveness of Word2Vec in clustering translated texts of Quranic verses, while also providing a visualization of the clustering results in the form of a two-dimensional projection to facilitate interpretation. With this approach, it is hoped that the process of clustering topics in the Quranic text will no longer be entirely dependent on manual searches or subjective assumptions, but can instead be supported by more objective and systematic data-based mapping.

II. Related Works

Studies on clustering or analyzing Quranic texts using Natural Language Processing (NLP) technology have been widely conducted in recent years. Various methods have been applied to extract thematic information from Quranic verses, both in Arabic and in translations into various languages.

Research was conducted using the k-means clustering method to group Quranic verses in Indonesian based on the similarity of keywords in which they appeared [11], [12]. They used a TF-IDF (Term Frequency–Inverse Document Frequency)-based approach to represent text features before the clustering process. The results showed that the verses could be grouped into several dominant themes with fairly stable clustering performance.

On the other hand, Word2Vec is used to cluster religious texts [13], [14]. They utilized Word2Vec to construct word vectors from a dataset of hadiths and then grouped these hadiths based on their semantic meaning using hierarchical clustering. The results of this study indicate that Word2Vec is effective in capturing semantic relationships between religious texts.

A study by T. Mikolov et al., the original developers of Word2Vec, demonstrated that this method is capable of representing words as fixed-dimensional numeric vectors containing semantic and syntactic relationships between words [5], [15], [16]. Word2Vec has been widely applied in various NLP contexts, including non-English language text corpora and religious texts.

Meanwhile, the Silhouette Score method, as a clustering evaluation metric, was introduced by P. J. Rousseeuw [8], and remains a standard measure for assessing the quality of clustering results based on the distance between data. This metric is used to measure how well data fits within its own cluster compared to other clusters.

Unlike previous studies that have largely used TF-IDF or hierarchical clustering, this study focuses on clustering Indonesian translations of Quranic verses using Word2Vec as the word embedding method and K-Means Clustering as the clustering algorithm, with performance evaluation using the Silhouette Score [7], [12]. To date, studies that specifically combine these three approaches in the context of Qur'an translation are still relatively rare, so this research is expected to be an initial reference in the development of a machine learning-based thematic interpretation system.

III. Research Methods

This research uses an unsupervised learning approach with a clustering method to group Indonesian translations of Quranic verses based on their similarity of meaning. The research was conducted systematically through several main stages, starting with data collection, preprocessing, word

embedding, clustering, evaluation, and analysis of the results [17], [18]. The detailed methodological stages are as follows:

A. Data Collection

The data used in this research consists of Indonesian translations of Quranic verses obtained from Kaggle, consisting of information on surah numbers, verse numbers, and the content of the verse translations.

B. Data Understanding

In this stage, several basic checks were conducted to understand the structure, quality, and main characteristics of the Quranic verse translation data [19]. The steps taken included:

- Examination of the number of columns and rows of data to determine the dimensions of the dataset and the data type of each column.
- Identification of missing values and ensuring there are no blank values in important columns.
- Analysis of the distribution of the number of verses per surah, to determine the distribution of the number of verses in each surah.
- Calculate the average text length per verse based on the number of words.

C. Data Preparation

Based on the results of the data understanding stage, a series of data improvements and transformations were performed to prepare it for the modeling process [20]. The following steps were performed:

- Case Folding: Converts all text to lowercase.
- Cleaning: Removing non-letter characters such as numbers, symbols, and punctuation.
- Stopword Removal: Removing common Indonesian words that have no significant meaning using a stopword list from NLTK and the Sastrawi library.
- Stemming: Converting each word to its base form using a stemmer from the Sastrawi library.
- Tokenizing: Separating sentences into clean word tokens.
- Creating new features, namely preprocessed tokens, which will be used in the embedding process.

D. Data Modeling

1) Word Embedding

The cleaned text data was converted into numeric vectors using the Word2Vec method. Word2Vec maps words with similar meanings into a 100-dimensional vector space [14], [21], [22], [23]. The parameters used include:

- vector_size: 100
- windows: 5
- min_count: 2
- seeds: 42

Next, each verse is represented as the average vector of its constituent words.

2) Clustering

After obtaining the vector representation of each verse, a clustering process is performed using the K-Means algorithm. The optimal number of clusters is found by testing several values of k (2 to 10), then selected based on the highest Silhouette Score value.

E. Model Evaluation

Evaluation is performed using the Silhouette Score as the primary metric for measuring the quality of clustering results [24]. Silhouette Score values range from -1 to 1, where values closer to 1 indicate that the data object is in the appropriate cluster.

F. Results Visualization

To facilitate the interpretation of clustering results:

- Dimensionality reduction is performed using PCA (Principal Component Analysis) to 2 dimensions.
- The reduction results are visualized in the form of a 2D scatterplot, with a different color for each cluster.

G. Interpreting Clustering Results

Interpretation is carried out by:

- Displaying several example verses from each cluster.
- Determining the 20 words closest to the centroid of each cluster based on the Word2Vec results using cosine similarity. These words represent the main topics in each cluster [25].

IV. RESULT AND DISCUSSION

A. Experimental Results

This research successfully implemented a series of systematic text data processing steps to cluster verses from Indonesian translations of the Quran based on semantic similarity. The developed pipeline consisted of several main stages: text preprocessing, word vector representation using Word2Vec, clustering using K-Means Clustering, and clustering quality evaluation using the Silhouette Score metric. The data used was obtained from the public platform Kaggle, containing a collection of Indonesian translations of the Quran, including information on the surah number, verse number, and the translation content of each verse.

The initial preprocessing stage involved several important processes, including case folding to standardize letters to lowercase, cleaning to remove non-alphabetic characters, stopword removal using a combination of NLTK and Sastrawi stopword lists, and stemming to return words to their basic form. Next, the text was divided into word tokens. This process is crucial to ensure that the data used in the Word2Vec model learning process is clean, consistent, and accurately represents the meaning of words in the context of the Quran.

After the preprocessing process was complete, the cleaned text data was converted into a numerical representation using the Word2Vec model. The model

parameters used were vector_size=100, window=5, min_count=2, and seed=42. These values were chosen based on previous research demonstrating the effectiveness of a similar configuration in modeling religious texts and holy books. Each verse was represented as the average of its word vectors. This representation allowed the Quranic verses to be converted into a numerical form that could be processed by clustering-based machine learning algorithms.

The next step was to cluster the verse vector data using the K-Means Clustering algorithm. This process was performed with varying numbers of clusters (k) ranging from 2 to 10. To determine the optimal number of clusters, the Silhouette Score metric was used, which measures how similar objects in one cluster are to objects in other clusters. The higher the Silhouette Score, the better the quality of the resulting clustering. Figure 1 and Table 1 show the Silhouette Score results for each k value

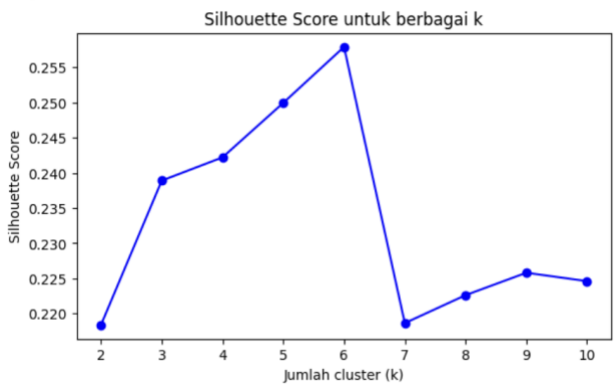

Fig 1. Plot silhouette score

Table 1. Silhouette Score

Number of Cluster (k)	Silhouette Score
2	0.2184
3	0.2389
4	0.2422
5	0.2500
6	**0.2579**
7	0.2186
8	0.2226
9	0.2258
10	0.2246

Based on these results, k=6 was chosen as the optimal number of clusters because it had the highest Silhouette Score, at 0.2579. This indicates that modeling with six clusters provides the best separation of verse groups based on similarity of meaning.

B. *Visualization of Clustering Results*

To obtain a visual representation of the distribution of the formed clusters, the clustering results were visualized in two dimensions using Principal Component Analysis (PCA). PCA was used to reduce the dimensionality from 100 to 2, allowing for a two-dimensional representation of the data.

This visualization in Gifure 2shows that the data is distributed into six clusters with different color distributions. Although there are areas of overlap, the six clusters are generally recognizable and well-separated. This indicates that Word2Vec successfully captures the semantic meaning between verses, and K-Means is able to map them into meaningful groups.

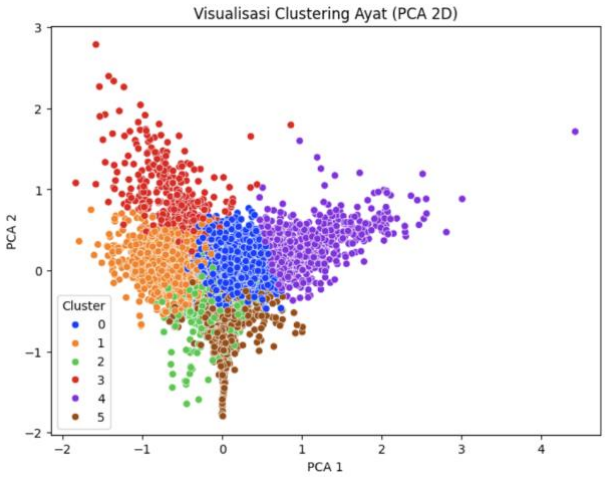

Fig 2. Clustering visualization

C. *Clustering Results Analysis*

To understand the characteristics of each cluster, we analyzed several example verses from each cluster and mapped the words closest to the cluster center (centroid) using cosine similarity [26]. These words reflect the dominant topic or central theme that shapes the semantic identity of each cluster. The analysis results show that each cluster has a fairly consistent and thematically relevant set of keywords, which can be grouped into specific categories based on the context of their meaning in the Quran. Table 2 presents the dominant topic words for each cluster in the Indonesian language.

Table 2. Dominant Word Topics of Each Cluster

Cluster	Dominant topic
0	*harap, putus, lepas, banyak, menang, izin, hak, mulia, zakat, doa, mu, serah, hadap, mudharat, liput, at, sumpah, buta* (hope, break up, let go, many, win, permission, right, noble, zakat, prayer, you, surrender, face, harm, cover, at, oath, blind)
1	*zalim, fasik, munafik, celaka, kafir, golong, anggap, yahudi, musyrik, rugi, nasrani, sombong, tambah, binasa, sebab, jihad, adakan, selamat, barangsiapa, sama* (unjust, wicked, hypocritical, wretched, infidel, class, considered, Jewish, idolatrous, loss, Christian, arrogant, add, perish, cause, jihad, organize, save, whoever, the same)
2	*sentuh, bapa, isterimu, tinggal, kawin, penuh, miskin, benda, suami, rumah, keji, yatim, budak, puasa, jenis, zina, kerabat, minum, khusus, pintu* (touch, father, your wife, stay, marry, full, poor, thing, husband, house, abominable, orphan, slave, fast, kind, adultery, relative, drink, special, door)
3	*terang, ajar, benar, taurat, kepada, baca, bukti, bahasa, tunjuk, rim, muhammad, wahyu, nyata, bawa, injil, jelas, ahli, ingkar, ingat, rasul* (light, teach, true, law, to, read, proof, language, point, ream, muhammad, revelation, real, bring, gospel, clear, expert, deny, remember, apostle)
4	*esa, murah, puji, karunia, santun, tuhan, suci, engkau, taubat, luas, punya, allah, raja, terima, ampun, kasih, dengar, kaya,*

Cluster	Dominant topic
	hendak, sungguh (one, cheap, praise, gift, polite, god, holy, you, repentance, vast, have, allah, king, accept, forgive, love, hear, rich, want, truly)
5	*sanggup, hilang, ucap, sulaiman, cela, dengan, atur, sulit, habis, ikan, laksana, sebut, wajah, atas, akan, pegang, susah, kaki, ganggu, langar* (able, lost, said, Solomon, blame, with, arrange, difficult, finished, fish, like, say, face, above, will, hold, difficult, feet, disturb, violate)

D. Analysis of Themes in Each Cluster

Cluster 0:
Dominated by words related to hope, prayer, charity, and a servant's inner attitude toward Allah. Words such as "harap," "doa," "zakat," "serah," and "waking" indicate a grouping of verses with themes of supplication and submission, as well as optimism regarding divine decree.

Cluster 1:
Contains words describing the concepts of moral evil, polytheism, disbelief, and the threat of punishment. Words such as "zalim," "fasik," "munafik," "kafir," "bisana," and "jihad" indicate that this cluster contains verses about threats to those who disbelieve, as well as calls for jihad and self-improvement.

Cluster 2:
Groups words related to social law, family, and interpersonal relationships. Vocabulary such as "wife," "husband," "marriage," "poor," "slave," "fasting," and "adultery" indicate that this cluster contains many verses about marriage laws, social obligations, and moral commands in social life.

Cluster 3:
Contains words that are closely related to the teachings of revelation, previous holy books, and previous prophets. Words such as taurah, muhammad, gospel, revelation, apostle, and read show that this cluster contains verses about the truth of revelation, the duties of the apostles, as well as warnings to previous people.

Cluster 4:
Raising the theme of divinity, the characteristics of God, and humans' inner relationship to Him. Words such as one, God, forgiveness, grace, repentance, and love indicate that this cluster contains verses that state the greatness of Allah, His merciful nature, and the call for humans to return to Him.

Cluster 5:
Contains words related to social behavior, character stories, and ethical commands in everyday life. Words such as able, said, fish, Solomon, woe, arrange, and hold show that the verses in this cluster contain many stories of wisdom and moral advice.

E. Interpretation of Findings

Based on the clustering and visualization results, it can be concluded that the Word2Vec and K-Means Clustering methods used in this study are capable of objectively identifying thematic groups in the translated text of the Quran. Each cluster successfully mapped a collection of verses with related meanings, reflected in the distribution of dominant words around the centroid of each cluster. These discovered semantic patterns offer significant potential for the development of various applications, such as a thematic verse search system, mapping relationships between themes in the Quran, a topic-based verse recommendation system, and the development of an automated topic-based learning system [27], [28], [29], [30], [31], [32] .

The success of this clustering demonstrates that despite the Quranic text's complex linguistic structure, distinctive style of delivery, and deep and contextual meaning, a word embedding-based approach can still be used to capture thematic meaning with considerable accuracy [30]. This provides initial evidence that natural language processing (NLP) technology can be integrated into Islamic studies, particularly in the context of developing data-driven digital interpretations and automated Qur'anic information retrieval [26]. Furthermore, the results of this study can also serve as a basis for developing a computer-based Qur'anic topic exploration system, which has previously been largely manual.

However, several interesting findings emerged during the clustering process, including verses that tend to be ambiguous or can be categorized into more than one theme. This overlap occurs because, in many cases, a verse contains more than one message context, or words within the verse have close semantic associations with several themes [33], [34], [35]. This presents both a challenge and an opportunity for further development, for example, by adding contextual information such as the asbabun nuzul (prophetic signs), makkiyyah or madaniyyah category metadata, and classical and contemporary interpretations to enrich the semantic context. Furthermore, static vector-based word embedding approaches such as Word2Vec still have limitations in capturing the contextual meaning between sentences or verses in their entirety. Therefore, future research could consider integrating transformer-based contextual embeddings such as IndoBERT to obtain richer meaning representations.

Overall, the results of this study provide a positive contribution to the use of NLP technology in Islamic studies. The automatic discovery of thematic patterns in the text of the Quran using a word embedding-based clustering method demonstrates that this approach is worthy of further development in the context of digitizing and automating data-driven exploration of Quranic content [25], [26], [27], [28], [29].

This study successfully applied the Word2Vec method as a word embedding technique and the K-Means Clustering algorithm to thematically group Indonesian translations of Quranic verses. Test results using the Silhouette Score demonstrated adequate clustering quality, while two-dimensional visualization using PCA provided a clear interpretation of the clustering results [36]. This approach demonstrates the potential of unsupervised learning to support the automation of thematic interpretation processes, which are currently performed manually by experts.

This research also shows a contribution in the development of technology-based systems for Islamic studies, and is in line with the achievement of Sustainable

Development Goals (SDGs) point 4, namely increasing access to inclusive and quality technology-based education. A limitation of this study is that the dataset used only includes Indonesian translations of Quranic verses without considering the context of interpretation or the historical background of the verses. Then, the representation of verse meaning is based solely on word distribution without considering the deeper semantic context or inter-verse relationships across surahs. The cluster quality assessment still relies on the Silhouette Score without validation from tafsir or Islamic scholars.

V. CONCLUSION

This study shows that the application of the Word2Vec method as a word embedding technique combined with the K-Means Clustering algorithm is able to group thematically translated texts of the Indonesian Quranic verses with quite good results, as indicated by the adequate Silhouette Score value and the two-dimensional PCA visualization that helps interpret the clustering results. This approach proves the potential of unsupervised learning in supporting the automation of the thematic interpretation process that was previously carried out manually, while contributing to the development of technology for Islamic studies and supporting the SDGs point 4 goal regarding inclusive and quality technology-based education. However, this study still has limitations, such as the use of a corpus that is limited to Indonesian translations, the lack of a contextual approach that considers the semantic and historical relationships of the verses, and an evaluation that still relies on quantitative metrics without expert validation. Therefore, further research is recommended to utilize more sophisticated embedding models such as BERT, integrate interpretation sources and verse metadata, involve interpretation experts in the evaluation process, expand the study to a multilingual corpus, and develop interactive web-based or mobile applications so that clustering results can be utilized more widely by the community.

REFERENCES

[1] S. M. Al-Qaththan, *Pengantar Studi Ilmu Al-Qur'an*. Pustaka Al-Kautsar, 2018.

[2] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, Sep. 1999, doi: 10.1145/331499.331504.

[3] A. Jaeger and D. Banks, "Cluster analysis: A modern statistical review," *WIREs Computational Statistics*, vol. 15, no. 3, May 2023, doi: 10.1002/wics.1597.

[4] G. J. Oyewole and G. A. Thopil, "Data clustering: application and trends," *Artif. Intell. Rev.*, vol. 56, no. 7, pp. 6439–6475, Jul. 2023, doi: 10.1007/s10462-022-10325-y.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Sep. 2013.

[6] K. Backhaus, B. Erichson, S. Gensler, R. Weiber, and T. Weiber, *Multivariate Analysis*. Wiesbaden: Springer Fachmedien Wiesbaden, 2021. doi: 10.1007/978-3-658-32589-3.

[7] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.

[8] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis Journal of Computational and Applied Mathematics 1987; 20: 53-65," *visited on*, pp. 4–13, 2023.

[9] United Nations, "Sustainable Development Goals (SDGs) 2030," sdgs.un.org. Accessed: May 12, 2025. [Online]. Available: https://sdgs.un.org

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Sep. 2013.

[11] I. Akbar, M. Faisal, and T. Chamidy, "Penerapan long short-term memory untuk klasifikasi multi-label terjemahan Al-Qur'an dalam Bahasa Indonesia," *JOINTECS (Journal of Information Technology and Computer Science)*, vol. 7, no. 1, pp. 41–54, 2024.

[12] R. M. Adani, P. P. Adikara, and N. Santoso, "Analisis Klaster Terjemahan Ayat Al-Qur'an Berbahasa Indonesia Menggunakan K-Means dan Word Embedding," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 9, no. 9, 2025.

[13] M. N. AD, I. Godbole, P. M. Kapparad, and S. Bhattacharjee, "Comparative analysis of religious texts: NLP approaches to the Bible, Quran, and Bhagavad Gita," in *Proceedings of the new horizons in computational linguistics for religious texts*, 2025, pp. 1–10.

[14] E. H. Mohamed and W. H. El-Behaidy, "An Ensemble Multi-label Themes-Based Classification for Holy Qur'an Verses Using Word2Vec Embedding," *Arab. J. Sci. Eng.*, vol. 46, no. 4, pp. 3519–3529, Apr. 2021, doi: 10.1007/s13369-020-05184-0.

[15] T. Mikolov, W. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 2013, pp. 746–751.

[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Adv. Neural Inf. Process. Syst.*, vol. 26, 2013.

[17] R. Brochier, A. Guille, and J. Velcin, "Global Vectors for Node Representations," in *The World Wide Web Conference*, New York, NY, USA: ACM, May 2019, pp. 2587–2593. doi: 10.1145/3308558.3313595.

[18] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Oct. 2018.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[21] S. Banerjee, "Word2Vec — a baby step in Deep Learning but a giant leap towards Natural Language Processing." [Online]. Available: https://medium.com/explore-artificial-intelligence/word2vec-a-baby-step-in-deep-learning-but-a-giant-leap-towards-natural-language-processing-40fe4e8602ba

[22] Md Al-Amin, M. S. Islam, and S. Das Uzzal, "Sentiment analysis of Bengali comments with Word2Vec and sentiment information of words," in *ECCE 2017 - International Conference on Electrical, Computer and Communication Engineering*, Institute of Electrical and Electronics Engineers Inc., Apr. 2017, pp. 186–190. doi: 10.1109/ECACE.2017.7912903.

[23] N. Rezki, S. A. Thamrin, and S. Siswanto, "Sentiment Analysis of Merdeka Belajar Kampus Merdeka Policy Using Support Vector Machine with Word2Vec," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 17, no. 1, pp. 0481–0486, Apr. 2023, doi: 10.30598/barekengvol17iss1pp0481-0486.

[24] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*, vol. 39. Cambridge University Press Cambridge, 2008.

[25] J. S. Coleman, *Introducing speech and language processing*. Cambridge university press, 2005.

[26] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.
[27] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of massive data sets*. Cambridge university press, 2020.
[28] R. Anand and U. Jeffrey David, *Mining of massive datasets*. Cambridge university press, 2011.
[29] J. D. Ullman, J. Leskovec, and A. Rajaraman, *Mining of massive datasets*, vol. 112. Cambridge University Press Cambridge, 2011.
[30] A. Ram, S. Jalal, A. S. Jalal, and M. Kumar, "A density based algorithm for discovering density varied clusters in large spatial databases," *Int. J. Comput. Appl.*, vol. 3, no. 6, pp. 1–4, 2010.
[31] M. Parimala, D. Lopez, and N. C. Senthilkumar, "A survey on density based clustering algorithms for mining large spatial databases," *International Journal of Advanced Science and Technology*, vol. 31, no. 1, pp. 59–66, 2011.
[32] H. Jiawei, M. Kamber, J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012. doi: 10.1016/B978-0-12-381479-1.00001-0.
[33] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE.," *Journal of machine learning research*, vol. 9, no. 11, 2008.
[34] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.
[35] A. Platzer, "Visualization of SNPs with t-SNE," *PLoS One*, vol. 8, no. 2, p. e56883, Feb. 2013, doi: 10.1371/journal.pone.0056883.
[36] T. Brown *et al.*, "Language models are few-shot learners," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 1877–1901, 2020.