J-KOMA Journal of Computer Science and Applications

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

Open Access

E-ISSN: *2620 - 4872*
Vol.*07*, ***No***.*02*
Doi: *https://doi.org/10.21009/j-koma.v7i2.05*

Recivied: *21 Okt 2024*
Accepted: *15 Nov 2024*
Published: *20 Des 2024*

Keywords:
Information Retrieval;
Word2Vec
Word Mover's Distance;
Cosine Similarity;
News Search.

Correspondence Email:
winangsari@uai.ac.id

Application of Information Retrieval in News Document Search Using Syntax and Semantic Orientation

Anggi Kurniawati[1], Winangsari Pradani[2*]
[1,2]Study Program of Informatics Engineering, Al Azhar University Indonesia.

Abstract
This study explores an information retrieval system for news document search, leveraging both syntactic and semantic approaches. The Word2Vec model, utilizing the skip-gram architecture, is employed to capture semantic relationships between words, transforming news articles into vector representations. Semantic similarity is measured using Word Mover's Distance (WMD) and Cosine Similarity, while a syntax-based method employs regular expressions for keyword matching. The dataset comprises 2,813 news articles from Liputan6.com and Tempo.co, collected between 25–31 August 2019, containing 25,951 unique words. Preprocessing steps include case folding, filtering, tokenization, stopword removal, and stemming to enhance data quality. The system was evaluated using six user queries, with performance assessed through Precision@k and Mean Average Precision (MAP). Results indicate that Word2Vec with Cosine Similarity achieved the highest MAP score of 76.86%, outperforming WMD (75.65%) and regular expressions (72.06%). This demonstrates the effectiveness of semantic-based retrieval for news documents. Future work should focus on larger datasets and advanced models like Doc2Vec to improve retrieval accuracy and contextual understanding.

INTRODUCTION

The rapid advancement of technology has significantly increased the number of internet users in Indonesia. According to the Indonesian Internet Service Providers Association (APJII), the number of internet users grew from 143.26 million in 2017 to 171.17 million in 2018, out of a total population of 264 million. More than half of Indonesia's population now relies on the internet as an integral part of daily life [1], [2], [3]. One of the main uses of the internet is information retrieval through search engines, which help users access relevant information quickly and effectively.

Search engines are information retrieval systems designed to return relevant results based on user queries. To perform this task, a search engine must be capable of identifying the similarity between user-entered keywords and existing documents. The importance of efficient information retrieval has grown alongside the rapid spread of online news, where users can easily access the latest updates through news portals such as Tribunnews, Detik, Okezone, Sindonews, Kompas, Liputan6, Kumparan, IDN Times, Merdeka, and Bolasport [4], [5].

Several approaches have been developed to enhance information retrieval performance. One of the notable methods is Word2Vec, introduced by Mikolov et al. at Google. Word2Vec provides a distributed representation of words and phrases by employing the skip-gram model, which is computationally efficient and capable of capturing semantic relationships between words. This model can represent semantic similarities and perform vector operations to generate more meaningful word associations. Furthermore, more recent approaches such as Word Mover's Distance (WMD) and Cosine Similarity have been introduced to measure semantic similarity between documents and queries more effectively.

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

Previous studies using Word2Vec have primarily focused on tasks such as text classification or recommendation systems. For instance, research on movie recommendation using synopses and titles combined with cosine similarity still faced challenges in achieving accurate results. However, the potential of Word2Vec for document similarity analysis, particularly in news retrieval, has not been extensively explored, as Word2Vec was only introduced in 2013 and WMD in 2016 [6], [7].

Based on these gaps, this study applies a semantic-based retrieval approach using Word2Vec combined with Word Mover's Distance and Cosine Similarity, alongside a syntax-based retrieval method using regular expressions. By comparing both approaches, the research aims to provide a more effective and context-aware document retrieval system for online news articles.

METHOD

Data Preprocessing

The data obtained from crawling must undergo preprocessing to reduce computational load, improve processing speed, and increase accuracy by extracting the essential terms. Preprocessing is the initial stage of text processing, which transforms unstructured documents into structured data for further analysis. This process generally involves several steps, including:

- **Case Folding**
  Converts all characters in the document into lowercase to achieve uniformity.
- **Filtering**
  Removes non-alphabetic characters, punctuation, and numbers, leaving only alphabetic tokens for analysis.
- **Tokenizing**
  Splits sentences into individual words using whitespace as the delimiter.
- **Stopword Removal**
  Eliminates common words that have little semantic value (e.g., "and," "from," "the"), allowing focus on more meaningful terms.
- **Stemming**
  Reduces words to their root form by removing affixes, ensuring that words with similar meanings are standardized.

Word2Vec Modeling

The tokenized data from the preprocessing stage is transformed into vector representations using the Word2Vec model [8], [9]. This study employs the skip-gram architecture implemented through the Gensim library, which effectively captures semantic relationships between words. Several parameters were set during training:

- Window size: defines the context range of surrounding words.
- Vector dimension: set to 150 for optimal representation.
- Epochs: training repeated seven times over the dataset.
- Minimum word frequency: ensures only words with sufficient occurrence are represented.

This process produces word embeddings capable of capturing both syntactic and semantic word similarities.

Semantic Similarity Computation

The similarity between user queries and documents is calculated using two semantic-based methods:

1. Word Mover's Distance (WMD)
   Measures the minimum "travel distance" required to align the distribution of words in a query with that in a document, leveraging word embeddings from Word2Vec [10].
2. Cosine Similarity

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

Computes the cosine of the angle between two document vectors. The closer the value is to 1, the more semantically similar the vectors. Cosine distance is defined as $1 - \text{cosine similarity}$.

Syntax-Based Retrieval with Regular Expression

In addition to semantic retrieval, a syntax-based approach is implemented using **regular expressions (regex)**. This method searches documents by matching exact keywords or combinations of keywords with logical operators (e.g., AND, OR).

Evaluation Method

The performance of the retrieval system is evaluated using:

1. **Precision@k**

   Evaluates the proportion of relevant documents among the top-k retrieved [11], [12], [13] documents

$$Precision@k = \frac{Number\ of\ relevant\ retrieved\ documents}{Total\ retrieved\ documents}$$

2. **Mean Average Precision (MAP)**

   Measures the average precision across all queries, providing an overall effectiveness score of the retrieval system [14], [15].

RESULTS & DISCUSSION

Data Collection

The dataset used in this study was collected from two Indonesian online news portals, Liputan6.com and Tempo.co. These portals were selected based on their popularity, as both are ranked among the top ten most-visited news websites in Indonesia. The data was gathered during the period of 25–31 August 2019, resulting in a total of 2,849 news articles (2,189 from Liputan6 and 660 from Tempo), with a total size of approximately 6.3 MB. Data extraction was performed using the Python Scrapy library.

Data Exploration

The collected articles were processed through a text preprocessing pipeline, including case folding, filtering, tokenization, stopword removal, and stemming. This step aimed to normalize the text and improve retrieval accuracy. From the total dataset of 2,849 articles, the documents were distributed across multiple categories with varying frequencies.

Table 1. Number of Articles in Each News Category

No	Category	Number of Articles
1	bisnis	347
2	bola	316
3	cantik	110
4	creativelab	2
…	…	…
21	tempo	1
22	travel	71

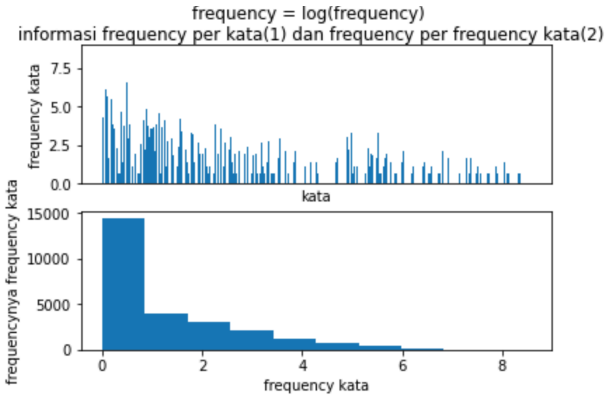
As shown in the table above, there are several outlier categories with only a small number of articles. Outliers are extreme data points that may disrupt distribution patterns and negatively affect the reliability of the research.

To address this issue, the outlier categories were removed. After eliminating these outliers, the dataset was reduced from 2,849 articles to 2,813 articles. Categories with fewer than 15 articles were excluded, including Tempo, Creativelab, Infografis, Rajut, Focus, and Kolom.

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

Furthermore, the news categories were modified. The Sport and Bola categories were merged into a single category, namely Olahraga (Sports), since “Bola” is inherently a subcategory of sports. This adjustment ensured a more consistent and representative dataset for analysis.

Thus, the final categories used in this study are: olahraga, metro, nasional, teknologi, bisnis, cantik, gaya, dunia, travel, otomotif, selebriti, politik, peristiwa, megapolitan, and berita.



Fugire 1. Word Frequency

In Figure 1, the word frequency distribution shows that out of a total of 25,951 unique words, more than 10,000 words appeared only once. Specifically, 11,038 unique words occurred a single time. Although this number is relatively large, in this study such words were retained, since the Word2Vec skip-gram model is able to represent rare words or phrases effectively. The average word frequency was 156, which was used as a reference in determining the embedding size during Word2Vec training. Words that appeared only once were also considered in setting the minimum word parameter for the skip-gram architecture.

Furthermore, to identify the most frequent words in each news category, the analysis revealed the highest-frequency terms per category, as summarized in Table 2.

Table 2. Word Frequency Distribution per News Category.

Sport		Travel		Automotive	
Word	**Frequency**	**Word**	**Frequency**	**Word**	**Frequency**
main	1643	wisata	161	mobil	324
liga	1132	kota	139	motor	238
tim	827	milik	107	listrik	209
laga	736	air	106	suzuki	192
latih	642	pulau	103	kendara	195

The majority of the words listed in Table 2 represent the most frequently occurring terms within each news category. For example, the word “main” appears most frequently in the olahraga category, reflecting its wide range of variations and relevance to sports news. Similarly, the word “wisata” is the most frequent term in the travel category, indicating its strong association and frequent usage in travel-related news. In the otomotif category, the word “mobil” appears most frequently, serving as a representative keyword with multiple variations commonly found in automotive news.

Word2Vec Training Results

The training process aimed to analyze the semantic similarity between word pairs using the Word2Vec model with different window sizes. The evaluation was carried out using two approaches. First, similarity was measured based on word pairs, where the average similarity score increased as the number of epochs progressed, both with a window size of 5 and 11.

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

Second, evaluation was performed using arithmetic operator combinations to examine how well the model could capture semantic relationships and represent words with similar meanings. The following section presents examples of similar words generated from several predefined keywords.

Table 3. Similar Words for "libur"

Word	Distance
cuti	0.8062671422958374
natal	0.7444887757301331
lotere	0.6768401861190796
elizabeth	0.6737784147262573
lancong	0.6722036004066467

Table 4. Similar Words for "kulit"

Word	Distance
jerawat	0.8054407835006714
pori	0.7841842770576477
sumbat	0.7817394733428955
kelupas	0.7780382633209229
iritasi	0.7725955843925476

In the evaluation of the Word2Vec method, the researcher determined that epoch 4 represented the average learning outcome when using epoch 7. During training with window size = 5, embedding size = 150, down-sampling = 1e-2, and epoch = 4, Word2Vec with 150 dimensions and a window size of 5 provided fairly good results in capturing semantic relationships within the data. This indicates that the combination of 150 dimensions and a window size of 5 is well-suited for representing the dataset. Conversely, when using smaller dimensions and larger window sizes, the resulting vectors fail to adequately capture the intended semantic meaning.

Results Using Word2Vec + Word Mover's Distance

Document retrieval using the semantic-based Word2Vec method combined with Word Mover's Distance (WMD) calculates semantic similarity between words. The results show that some documents are considered relevant or irrelevant depending on the query entered. The retrieval process is then evaluated based on whether the returned documents contain semantic relevance during the search. The following presents the results of document retrieval using the Word2Vec and Word Mover's Distance methods.

Table 5. Document Retrieval Results Using Word2Vec + WMD

Query	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8	Doc 9	Doc 10
strategi dagang	1	0	1	1	1	1	1	1	0	1
krisis ekonomi	0	0	0	1	1	1	1	1	0	1
turnamen esports indonesia	0	0	1	0	1	0	0	0	0	0
medali badminton indonesia	1	1	1	1	1	1	1	1	1	1
tuan rumah piala dunia	1	1	1	0	0	0	0	0	1	0
masalah kulit kering	1	1	1	1	1	1	1	1	1	1

Results Using Word2Vec and Cosine Distance

Document retrieval using the semantic word2vec method with similarity calculations based on cosine distance shows that several documents were considered relevant or irrelevant depending on the query submitted. The evaluation process was carried out by assessing whether the retrieved documents contained semantic similarity to the query keywords.

The results presented in Table 4.6 indicate that the combination of word2vec with cosine distance is able to identify relevant documents for most queries, although the number of relevant documents varies

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

across different cases. For instance, for the query “strategi dagang” (trade strategy), most documents were considered relevant with a value of 1, whereas for the query “krisis ekonomi” (economic crisis), only a few documents were retrieved as relevant. This suggests that the performance of retrieval is highly dependent on the semantic closeness between words in the vector space produced by word2vec.

In general, this approach demonstrates reasonably good performance in capturing semantic relationships between documents. However, it is not as optimal as the word2vec method combined with Word Mover’s Distance, which yielded higher overall relevance in document retrieval.

Results Using Regular Expression

Document retrieval using the syntactic method regular expression shows several documents considered relevant and irrelevant depending on the query entered. The retrieval process was then evaluated based on whether the output contained semantic similarity or not during the search. The following is the result of document retrieval using the syntactic method based on regular expression.

Table 6. Document results using regex

Query	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8	Doc 9	Doc 10
strategi dagang	1	0	0	0	1	1	1	1	0	0
krisis ekonomi	1	0	1	0	0	1	1	0	1	0
turnamen esports indonesia	1	1	1	1	0	0	0	0	0	0
medali badminton indonesia	1	1	1	0	1	0	0	0	0	0
tuan rumah piala dunia	1	1	0	1	0	1	0	0	0	0
masalah kulit kering	1	0	1	0	1	0	0	0	0	0

Evaluation

This study used 2,813 news articles, resulting in a total of 25,951 unique words. The evaluation process applied each query to assess whether the retrieved news documents were relevant or irrelevant to the query.

Table 7. Document results using regex

Method	Relevant Documents	Irrelevant Documents
Word2Vec + Word Mover's Distance	40	20
Word2Vec + Cosine Similarity	34	26
Regular Expression	24	36

Based on Table 4.8, the results show that the Word2Vec combined with Word Mover’s Distance method achieved the highest number of relevant documents, with 40 relevant out of six queries. The Word2Vec combined with Cosine Similarity can also be considered fairly effective, retrieving 34 relevant documents. Meanwhile, the Regular Expression method retrieved 24 relevant documents.

Evaluation Using precision@k

The evaluation performed in this study also measured the value of *precision@k* with k = 10. The top 10 retrieved news documents were taken as a reference for evaluation. The precision@k value was then calculated for each query.

As shown, the Word2Vec + Word Mover’s Distance method achieved average precision values of 60.81%, 49.08%, 36.50%, 100%, 86.10%, and 100% for the first six queries, respectively. The Word2Vec + Cosine Similarity method obtained 96.16%, 30.75%, 45.70%, 88.56%, and 100%. The Regular Expression method obtained 36.25%, 65.60%, 100%, 91.50%, 85.25%, and 75.23%.

Table 8. precision@k results

Query	Word2Vec + WMD	Word2Vec + Cosine Similarity	Regular Expression
strategi dagang	60.81%	96.16%	36.25%
krisis ekonomi	49.08%	30.75%	65.60%
tuan rumah piala dunia	36.50%	45.70%	100%

turnamen esports indonesia	100%	88.56%	91.50%
medali badminton indonesia	86.10%	100%	85.25%
masalah kulit kering	100%	85.25%	75.33%

Evaluation Using Mean Average Precision (MAP)

Another evaluation metric used was Mean Average Precision (MAP). This metric measures the average retrieval performance of each method based on multiple queries.

Table 9. Mean Average Precision (MAP) results

Method	MAP (%)
Word2Vec + WMD	75.65%
Word2Vec + Cosine Similarity	76.86%
Regular Expression	72.06%

Table 9 shows that Word2Vec + Cosine Similarity achieved the best MAP value of 76.86%, slightly higher than Word2Vec + WMD (75.65%) and Regular Expression (72.06%). These results indicate that Cosine Similarity provided better retrieval accuracy overall

CONCLUSION

The results of this study indicate that the Word2Vec model with an embedding size of 150, window size of 5, minimum word count of 1, and four training epochs provided the best performance for the given dataset. The implementation of Word2Vec combined with Word Mover's Distance and Cosine Similarity methods showed strong capability in retrieving semantically similar news documents, achieving an overall accuracy above 70%. Among the tested methods, Word2Vec with Cosine Similarity achieved the highest Mean Average Precision (MAP) score of 76.86%, demonstrating its effectiveness in semantic-based document retrieval. Future research is recommended to enhance model generalization by training on larger and more diverse corpora, experimenting with deeper architectures, and comparing with other semantic models such as Doc2Vec to further improve retrieval accuracy and contextual understanding.

REFERENCES

[1] A. Siregar, D. Trirahayu, and A. Achmadi, "The Review of Financial Technology Services in Indonesia," in *First International Conference on Humanities, Education, Language and Culture, ICHELAC 2021, 30-31 August 2021, Flores, Indonesia*, EAI, 2021. doi: 10.4108/eai.30-7-2021.2313965.

[2] D. K. Sari, Kartika Ayu Rachmawati, Byba Melda Suhita, Lingga Kusumawardani, and Dedi Saifulah, "The Influence of Social Media Addiction on Adolescent Self-Concept," *Journal Of Nursing Practice*, vol. 7, no. 1, pp. 39–44, Oct. 2023, doi: 10.30994/jnp.v7i1.428.

[3] M. Pratiwi P, E. Rosnawati, M. T. Multazam, and N. F. Mediawati, "Personal Data Collection: Recent Developments in Indonesia," *KnE Social Sciences*, Aug. 2022, doi: 10.18502/kss.v7i12.11503.

[4] Z. Qathrunnada, C. Nugroho, F. Yusanto, A. Wulandari, and R. R. Wulan, "Ideology, resistance, and sociopolitical dynamics in Indonesia: media narratives and resistance discourses on the chairman of the corruption eradication commission's corruption case," *Front Commun (Lausanne)*, vol. 10, Feb. 2025, doi: 10.3389/fcomm.2025.1552110.

[5] D. Subekti, M. Yusuf, M. Saadah, and M. Wahid, "Social media and disinformation for candidates: the evidence in the 2024 Indonesian presidential election," *Front Polit Sci*, vol. 7, Jul. 2025, doi: 10.3389/fpos.2025.1625535.

[6] M. B. Dehkordi, A. Zaraki, and R. Setchi, "Optimal Feature Set for Smartphone-based Activity Recognition," *Procedia Comput Sci*, vol. 192, pp. 3497–3506, 2021, doi: 10.1016/j.procs.2021.09.123.

[7] G. Zheng, J. Yao, Z. Tu, and X. zhou, "Effect of Building Height on Wind Load Characteristics of Photovoltaic Arrays," *J Phys Conf Ser*, vol. 2399, no. 1, p. 012003, Dec. 2022, doi: 10.1088/1742-6596/2399/1/012003.

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

[8] D. E. Cahyani and I. Patasik, "Performance comparison of TF-IDF and Word2Vec models for emotion text classification," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2780–2788, Oct. 2021, doi: 10.11591/eei.v10i5.3157.

[9] A. Sharma and S. Kumar, "Ontology-based semantic retrieval of documents using Word2vec model," *Data Knowl Eng*, vol. 144, p. 102110, Mar. 2023, doi: 10.1016/j.datak.2022.102110.

[10] T. A. Adjuik and D. Ananey-Obiri, "Word2vec neural model-based technique to generate protein vectors for combating COVID-19: a machine learning approach," *International Journal of Information Technology*, vol. 14, no. 7, pp. 3291–3299, Dec. 2022, doi: 10.1007/s41870-022-00949-2.

[11] S. Airen and J. Agrawal, "Movie Recommender System Using K-Nearest Neighbors Variants," *National Academy Science Letters*, vol. 45, no. 1, pp. 75–82, Feb. 2022, doi: 10.1007/s40009-021-01051-0.

[12] D. K. Waweru, F. Yang, C. Zhao, and L. M. Paul, "P-Adic LDPC Codes at Precision k," *Wirel Pers Commun*, vol. 139, no. 2, pp. 1269–1283, Nov. 2024, doi: 10.1007/s11277-024-11667-2.

[13] M. H. Joseph and S. D. Ravana, "Improving the accuracy of the information retrieval evaluation process by considering unjudged document lists from the relevant judgment sets," *Information Research an international electronic journal*, vol. 29, no. 3, Sep. 2024, doi: 10.47989/ir293603.

[14] A. M. Roy and J. Bhaduri, "Real-time growth stage detection model for high degree of occultation using DenseNet-fused YOLOv4," *Comput Electron Agric*, vol. 193, p. 106694, Feb. 2022, doi: 10.1016/j.compag.2022.106694.

[15] M. Chen *et al.*, "Improved faster R-CNN for fabric defect detection based on Gabor filter with Genetic Algorithm optimization," *Comput Ind*, vol. 134, p. 103551, Jan. 2022, doi: 10.1016/j.compind.2021.103551.