

Exploring Semantic Similarity among MUI Fatwas: A Computational Analysis using Generalized Jaccard Similarity

Alfado Rafly Hermawan^{a*}, Shofa Wardatul Jannah^a

^a Information System, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya, 61211, Indonesia.

ABSTRACT

Majelis Ulama Indonesia (MUI) plays a crucial role in the Islamic landscape of Indonesia, influencing religious discourse and societal norms. As a primary contributor to policy formulation and the issuance of Islamic fatwas, the MUI significantly impacts the lives of Muslims. However, challenges arise when certain fatwas exhibit similarities, necessitating deeper analysis to understand their differences. Despite limited prior research, there is an urgent need for a computational framework to comprehensively assess fatwa similarities. This study addresses this gap by employing the Generalized Jaccard Similarity method with WordNet, demonstrating its effectiveness compared to the Jaccard method with a 25.86% improvement in string matching quality for evaluating MUI fatwa titles. The Generalized Jaccard similarity analysis reveals that 73 documents exhibit similarity scores ≥ 0.5 , indicating significant resemblance, while 77,028 documents have scores < 0.5 , indicating lower similarity or dissimilarity. These figures reflect varying degrees of document similarity based on Generalized Jaccard.

Keywords: Generalized Jaccard Similarity, WordNet, Natural Language Processing, MUI, Islamic Documents.

© 2024 Pusat Kajian Halal ITS. All rights reserved.

1 Introduction

The Indonesian Ulama Council (Majelis Ulama Indonesia or MUI) is a prominent Islamic organization in Indonesia that operates as a pivotal institution within the intricate landscape of the country. MUI plays a central role in shaping religious discourse and guidance. This council, no stranger to debate over its pronouncements, fulfils a dual role: supporting the government in formulating social policies and serving as the leading Islamic fatwa [1]. It convenes representatives from diverse Islamic organizations and issues fatwas that significantly influence the lives of Muslims in Indonesia, serving as references for living in accordance with Islamic teachings [2].

Fatwas (Islamic documents), religious rulings issued by the MUI, encompass more than just content. These fatwas also address the method and purpose behind the ruling, ensuring a strong foundation and rationale to encourage public adherence based on the test of ijtihad [3]. As time progresses and the Muslim community encounters new complexities, MUI continues to issue fresh fatwas, leading to a vast collection covering a wide range of topics

* Corresponding author. Tel. +62 881-9845-544
Email address: 6026231041@student.its.ac.id

that extend beyond religious domains such as faith and worship to various aspects of human life.

However, a challenge arises when instances of similarity occur within certain MUI fatwas, as evidenced by cases such as the fatwa issued on February 1, 2010, concerning Qibla, which resembles another fatwa issued on July 26, 2010, addressing the same issue Qibla direction. To understand this perspective, a closer analysis of the fatwas themselves is necessary. Therefore, it is essential to conduct in-depth analysis to understand the similarities among these fatwas.

There has been a no research concerning the exploration of similarities among the fatwas issued by the MUI. This research gap underscores the necessity for the development of a computational framework specifically designed to evaluate the similarities among these diverse fatwas. To address this gap, the development of a specialized computational framework for evaluating the diverse fatwas is proposed. Specifically, the similarity between fatwas will be measured using the Generalized Jaccard Similarity method with WordNet. By adopting this approach, deeper insights into the semantic synonymic relationships among various fatwas issued by the MUI can be gained. Leveraging NLP techniques and methodologies, valuable information will be provided, enabling more efficient and accurate analysis for both researchers and practitioners.

2 Related Works

There are some studies relating to the similarity between documents and the application: Table 1. Summary of the related work.

Author	Pre-processing methods	Method	Dataset
Valentino Rossi et al. [4]	Removal Punctuation, WordNet	Agglomerative Hierarchical Clustering	The Indonesian Thesaurus
Nur Aini Rakhmawati et al. [5]	Remove duplicate, Lowercase, Punctuation Removal, Stopword, Tokenization	WordNet, Jaccard Similarity, Levenshtein Distance, dan Jaro-Winkler Distance	271, 823 Food Ingredients from Open Foods Fact and LODHalal
R. Gnanakumari et al. [6]	-	Deep Learning (Recurrent DNN - The Generalized Jaccard Suspicious Behaviour Similarity)	1,600,000 tweet and 250 tweet users from the dataset to identify the Suspicious Behaviours (SBs) of users
Raiffudin M [7]	-	TF-IDF, Word-Net, Wu-Palmer	Quran Sahih International and An-Nawawi Forty Hadith
Olowolayemoo A et al. [8]	-	TF-IDF, Multinomial Naive Bayes,	Islamic websites (Sunni's websites and Shia's websites)

Despite the scarcity of research on these methods, the Generalized Jaccard method to Islamic documents offers several advantages. For instance, it tolerates token variations, provides flexibility by combining set-based and sequence-based techniques, and enhances string matching quality compared to traditional string matching measurements [4]. In addition, discussions on document similarity in Islamic documents still primarily rely on commonly used approaches such as Jaccard Similarity, Levenshtein Distance, and Jaro-Winkler Distance [5]. Furthermore, the use of WordNet as a pre-processing step has proven effective in determining similarity between the same words in different documents [6]. Therefore, in this section, a solution is proposed to address the issue of determining similarity between Islamic documents, particularly MUI fatwas, using the Generalized Jaccard method and leveraging WordNet as an approach to obtain word similarity. Overall, the use of Generalized Jaccard Similarity is theoretically grounded to measure synonym similarity among MUI fatwas. This method is also capable of handling document length variations, synonyms, and word variations within the documents [7].

3 Background Theory

The background theory section, relevant literature sources supporting this research are discussed. In particular, concepts related to Jaccard similarity, Wordnet, and the Generalized Jaccard method are explored.

3.1 Jaccard Similarity

The Jaccard similarity is defined as the ratio of the number of items co-rated by two users to the number of items rated by at least one of them [8], [9] expressed as follows:

$$Jaccard(a, b) = \frac{|I_a \cap I_b|}{|I_a \cup I_b|} \quad (1)$$

- $Jaccard(a, b)$ = Represents the Jaccard similarity score between sets A and B.
- $|I_a \cap I_b|$ = Represents the number of elements that are in both set A and set B (intersection).
- $|I_a \cup I_b|$ = Represents the number of elements that are in either set A or set B (union).

3.2 Wordnet

WordNet is a lexical database that organizes words into categories based on their part of speech (noun, verb, adjective, etc.). Within each category, words are grouped into sets called synsets. These synsets represent groups of words with related meanings.

3.3 Generalized Jaccard Similarity

Generalized Jaccard Similarity (GJS) is a technique utilized in natural language processing (NLP) to gauge the similarity between sets of synonyms extracted from WordNet. The Jaccard similarity coefficient, used to compare similarities and differences between finite sample sets. A higher Jaccard score indicates greater similarity between the sets. The generalized Jaccard coefficient, also known as the Tanimoto coefficient [12]

GJS is computed using the following formula:

$$GJ(x, y) = \frac{\sum(x_i, y_j \in M^{s(x_i, y_j)})}{|B_x| + |B_y| - |M|} \quad (2)$$

Explanation:

- $\sum(X_i, Y_j) \in M^{s(X_i, Y_j)}$ = Represents the sum of the weights of common word pairs between the two documents.
- $|B_x|$ = Represent the total number of words in documents 1
- $|B_y|$ = Represent the total number of words in documents 2
- $|M|$ = Represents the number of common words between the two documents.

4 Methods

This section demonstrates the design that will be implemented in this research which can be seen in Figure 1. The first step is collecting dataset from official MUI website.

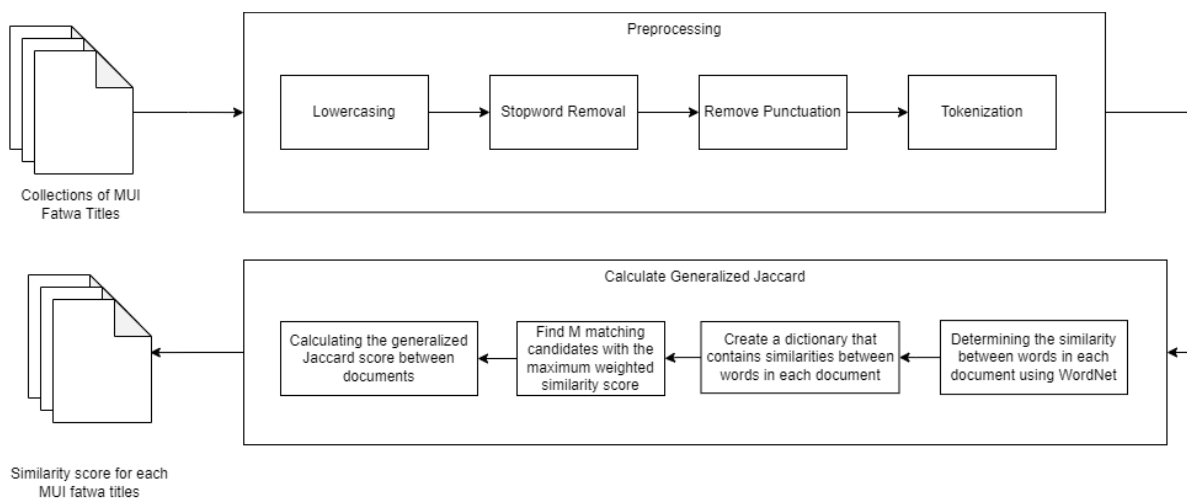


Fig. 1 Illustrative of overall research flow

4.1 Data Collection

The MUI fatwa dataset is sourced directly from the official MUI website. It is compiled through manual collection, employing a web scraping extension to automate the extraction of fatwa texts from the website. The dataset is structured in a Comma-Separated Values (CSV) format, containing 5 key attributes and the total number of scraped fatwas amounts to 393 fatwas.

Table 2. The raw dataset from web scraping

No	Fatwa Title	Fatwa Theme	Fatwa Number	Date Established
1	PANDUAN PENYELENGGARAAN IBADAH DI BULAN RAMADAN DAN SYAWAL 1442 H	Sosial Kemasyarakatan	KEPUTUSAN IJTIMA ULAMA KOMISI FATWA SE-INDONESIA TAHUN 2003	16-Dec-03

No	Fatwa Title	Fatwa Theme	Fatwa Number	Date Established
2	PENGUNAAN MIKROBA DAN PRODUK MIKROBIAL DALAM PRODUK PANGAN	POM Iptek	01 Tahun 2010	19-Jan-10

4.2 Pre-processing

Pre-processing is utilized to mitigate issues stemming from incomplete, interrupted, or inconsistent data Fatwas. critical stage in text analysis that aims to standardize and clean textual data to enhance its suitability for further analysis tasks. In this study, we employ several pre-processing techniques to mitigate issues stemming from incomplete, interrupted, or inconsistent data Fatwas for meaningful analysis.

Table 3. Fatwa Before Pre-processing

Before Pre-processing
PANDUAN PENYELENGGARAAN IBADAH DI BULAN RAMADAN DAN SYAWAL 1442 H
CARA PENSUCIAN EKSTRAK RAGI (YEAST EKTRACT) DARI SISA PENGOLAHAN BIR (BREWER YEAST)

- a. **Lowercasing:** This process involves converting all letters in the text to lowercase [13]. By doing so, consistency in word treatment is ensured regardless of their original casing. For example, "Hello" and "hello" would be considered the same word after lowercasing.

Table 4. Fatwa After Lowercasing

After Lowercasing
panduan penyelenggaraan ibadah di bulan ramadan dan syawal 1442 h
cara pensucian ekstrak ragi (yeast ekstract) dari sisa pengolahan bir (brewer yeast)

- b. **Punctuation Removal:** Punctuation marks such as commas, periods, question marks, and exclamation points are removed from the text. Punctuation marks often do not carry semantic meaning in text analysis tasks and their presence can sometimes interfere with downstream processing steps [14].

Table 5. Fatwa After Punctuation Removal

After Punctuation Removal
panduan penyelenggaraan ibadah di bulan ramadan dan syawal 1442 h
cara pensucian ekstrak ragi yeast ekstract dari sisa pengolahan bir brewer yeast

- c. **Stopwords Removal:** Stopwords are common words in a language that typically do not contribute much to the overall meaning of the text and do not affect to the classification process [15]. Examples of stopwords in Indonesia include "dan", "di", "yang", "ini", etc. Removing stopwords can help reduce the dimensionality of the data and focus on more meaningful terms. In the process of removing stop words, the Sastrawi library is used. The Sastrawi library is a Python library that makes it possible to trim words that are generally not used in analysis.

Table 6. Fatwa After Stopwords Removal

After Stopwords Removal
panduan penyelenggaraan ibadah bulan ramadan syawal 1442
cara pensucian ekstrak ragi yeast ekstrak dari sisa pengolahan bir brewer yeast

- d. Tokenization: Tokenization involves breaking down the text into smaller units, such as words or subwords [16]. In the context of fatwa texts, tokenization would divide the text into individual words. This step is essential for further analysis as it allows us to treat each word as a separate entity and apply operations or analyses at the word level.

Table 7. Fatwa After Punctuation Removal

After Punctuation Removal
panduan penyelenggaraan ibadah bulan ramadan syawal 1442
cara pensucian ekstrak ragi yeast ekstrakt sisa pengolahan bir brewer yeast

4.3 Generalized Jaccard Similarity Measurement

At this point, we begin by constructing a dictionary of synonyms, which includes a range of synonymous terms for the words identified in the title of the fatwa, utilizing the Indonesian WordNet. Subsequently, a dictionary is developed that assigns a similarity rating to each term across the documents of fatwa titles. Words that are identical or have synonymous counterparts are given a full similarity score of one, whereas words that differ between the documents are assigned a score of zero. Utilizing this dictionary of similarity, the Generalized Jaccard similarity is then computed, applying Equation (2). This section will illustrate a straightforward example of how to calculate the Generalized Jaccard similarity.

Example:

Document 1: ['pencurian', 'energi', 'listrik']

Document 2: ['aliran', 'ahmadiyah']

From the two documents above, the word 'listrik' is present in Document 1, while 'aliran' is present in Document 2. Both words are similar because they are related to flow and are synonyms, although in different contexts. Therefore, 'listrik' and 'aliran' have a similarity value. Here is an example calculation from the two documents:

$$\frac{1}{3 + 2 - 1} = 0.25$$

In this example, the words 'listrik' and 'aliran' have a semantic similarity because they both relate to the concept of flow, albeit in different contexts.

5 Results and discussions

In a dataset of 393 fatwas, a similarity measurement process was conducted to determine how similar each fatwa is to the others. This involved calculating the similarity score between each pair of fatwas using a method called Generalized Jaccard Similarity and Jaccard similarity. This method not only enables us to assess the degree to which two fatwas share similar content and themes, but it also facilitates the evaluation of how the Generalized Jaccard similarity can enhance the comparison of document similarities. The results of this measurement can provide valuable insights into the similarities and differences between the fatwas.

Table 8. The similarity results between fatwas with Generalized Jaccard Similarity scores

Fatwa 1	Fatwa 2	Similarity GJS	Similarity JS
['fatwa', 'penyerangan', 'amerika', 'serikat', 'sekutunya', 'irak']	['penyerangan', 'amerika', 'serikat', 'sekutunya', 'irak']	0.83	0.83
['pandangan', 'ruu', 'larangan', 'minuman', 'beralkohol']	['ruu', 'larangan', 'minuman', 'beralkohol']	0.8	0.8
['pelaksanaan', 'shalat', 'jumat', 'dzikir', 'kegiatan', 'keagamaan', 'tempat', 'masjid']	['pelaksanaan', 'shalat', 'jumat', 'dzikir', 'kegiatan', 'keagamaan', 'tempat', 'masjid']	0.78	0.78
['hukum', 'alkohol', 'minuman']	['hukum', 'alkohol']	0.67	0.67
['pencurian', 'energi', 'listrik']	['aliran', 'ahmadiyah']	0.25	0
['panduan', 'penyelenggaraan', 'ibadah', 'bulan', 'ramadan', 'syawal', '1442']	['arah', 'kiblat']	0.125	0

In an example regarding views on the bill prohibiting alcoholic beverages, there are two fatwas that share similarities but also differences.

Fatwa 1:

- Tokens: ['pandangan', 'ruu', 'larangan', 'minuman', 'beralkohol']
- Date issued: 09 May 2018

Fatwa 2:

- Tokens: ['ruu', 'larangan', 'minuman', 'beralkohol']

- Date issued: 09 June 2015

The two fatwas share the common ground of discussing the draft bill on banning alcoholic beverages. However, they may differ in their focus or scope of discussion. Additionally, the difference in the dates of issuance of the two fatwas suggests that the discussion of the bill has undergone development or changes over time. This could include revisions or additions to the content of the bill, changes in opinions or views on the bill, or possibly new developments that have influenced the discussion of the draft bill on banning alcoholic beverages.

Fig 2 shows the similarity scores between fatwas in a dataset revealing an overall similarity process score of 77.028. Individual scores range between 0.0 and 1.0, where 0.0 indicates no similarity and 1.0 signifies identical content. The distribution shows a significant number of fatwas with a similarity score of 0.0, indicating a diverse range of topics and content within the dataset. This diversity underscores the varied nature of the fatwas analysed. Importantly, using the Generalized Jaccard Measure (GJM) improves scores compared to the standard Jaccard measure, suggesting GJM's effectiveness in handling document variations and enhancing similarity assessments.

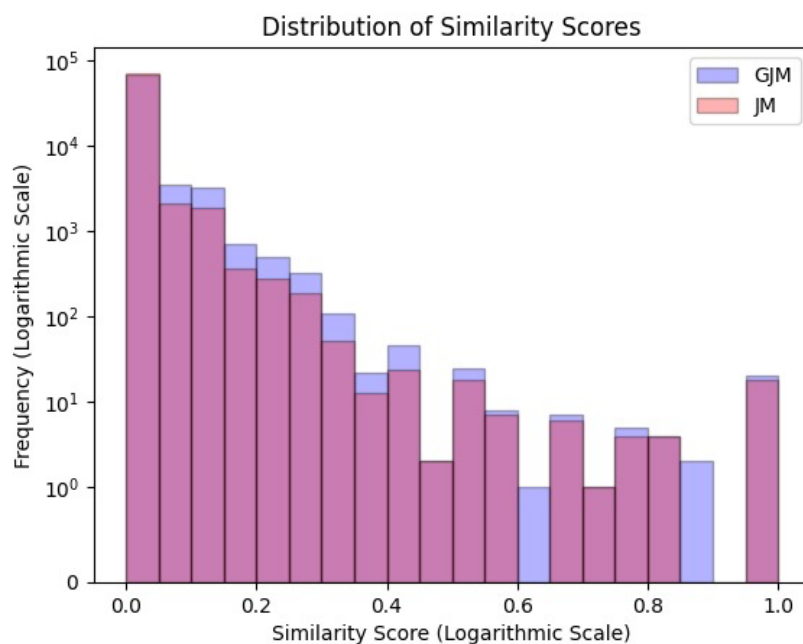


Fig. 1 All distribution similarity score

Fig 3 presents the distribution of similarity scores among fatwas in the dataset, ranging from 0.5 to 1.0, where higher scores indicate greater content overlap without being identical. The prevalence of scores between 0.5 and 1.0 suggests significant thematic similarities across fatwas, providing insights into recurring topics within the dataset. Specifically, the Generalized Jaccard Measure (GJM) identifies 73 fatwas with scores above 0.5 and 20 fatwas with a score of 1.0, demonstrating its ability to capture nuanced similarities despite variations in document attributes. In comparison, the Standard Jaccard Measure (JM) identifies 58 fatwas above 0.5 and 18 fatwas with a score of 1.0. Additionally, 77,028 fatwas exhibit scores below 0.5, indicating diverse content, while 68,396 fatwas have a score of 0.0, indicating no

6 Conclusion

This research aimed to measure the similarity between fatwas using the Generalized Jaccard Similarity method with WordNet. The study analysed a dataset of 393 fatwas, employing the Generalized Jaccard Similarity (GJM) method augmented with WordNet to assess similarity between each pair of fatwas. This approach effectively evaluated content and thematic overlap, demonstrating GJM's superiority over the standard Jaccard measure in comparing fatwas.

The findings revealed a nuanced spectrum of similarities and distinctions among the fatwas. For instance, analysis of a fatwa concerning a bill prohibiting alcoholic beverages highlighted shared core themes with potential variations in focus or perspective. The study also underscored the dataset's diversity, with numerous fatwas exhibiting no similarity, indicative of the broad range of topics covered. Moreover, a significant proportion of fatwas demonstrated high similarity values (between 0.5 and 1.0), suggesting thematic convergence and potential redundancy in established rulings.

Examining word frequencies in fatwa titles provided further insights into the issues addressed. Terms such as "hukum" (law), "zakat" (almsgiving), and "sholat" (prayer) reflected the predominant focus on Islamic jurisprudence and practice. Additionally, terms like "vaksin" (vaccine) and "covid" underscored the fatwas' engagement with contemporary concerns.

In conclusion, this study successfully employed GJM with WordNet to analyse fatwa similarity, offering valuable insights into their thematic landscape. The findings contribute to understanding the diversity and overlap of fatwa topics, laying the groundwork for future research on their categorization and organization to enhance accessibility and comprehension.

References

- [1] R. Muhaimin and J. Muslimin, "The Role of the Council of Indonesian Ulama (MUI) to the Development of a Madani Society in the Democratic Landscape of Indonesia," *Aspirasi: Jurnal Masalah-masalah Sosial*, vol. 14, no. 2, Dec. 2023, doi: 10.46807/aspirasi.v14i2.3368.
- [2] D. Khairani, A. Lubis, Zulkifli, H. T. Sukmana, K. Faruqi, and Y. Durachman, "Keywords Searching XML Data for Fatwa in Indonesia," in *2019 7th International Conference on Cyber and IT Service Management (CITSM)*, IEEE, Nov. 2019, pp. 1–4. doi: 10.1109/CITSM47753.2019.8965394.
- [3] I. Izmuiddin, R. Rusyaida, G. Bashir, P. Hasibuan, and A. Awaluddin, "The Indonesian Ulema Council Fatwa Analysis on The Environment and Their Relationship to The Green Economics Concept Development," *Jurnal Syntax Admiration*, vol. 3, no. 12, pp. 1559–1568, Dec. 2022, doi: 10.46799/jsa.v3i12.536.
- [4] V. R. Fierdaus, M. A. Bijaksana, and W. Astuti, "Building Synonym Set for Indonesian WordNet using Commutative Method and Hierarchical Clustering," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 4, no. 3, p. 778, Jul. 2020, doi: 10.30865/mib.v4i3.2254.

- [5] N. A. Rakhmawati and M. Jannah, "Food Ingredients Similarity Based on Conceptual and Textual Similarity," *Halal Research Journal*, vol. 1, no. 2, pp. 87–95, Oct. 2021, doi: 10.12962/j22759970.v1i2.107.
- [6] R. Gnanakumari and P. Vijayalakshmi, "Generalized Jaccard Similarity Based Recurrent DNN for Virtualizing Social Network Communities," *Intelligent Automation & Soft Computing*, vol. 36, no. 3, pp. 2719–2730, 2023, doi: 10.32604/iasc.2023.034145.
- [7] M. Raffiudin, "Exploring Sacred Texts: Leveraging Computer Science for Dataset Similarity Analysis in Religious Studies," Apr. 2024, pp. 227–235. doi: 10.4028/p-kE3XmS.
- [8] A. Olowolayemo, N. H. Daud, M. G. Tanni, and M. A. Omar Ba Khadher, "An All-Inclusive Digital Framework for Collaborative Community Transformation for Sustainable Development," *International Journal on Perceptive and Cognitive Computing*, vol. 9, no. 1, pp. 1–13, Jan. 2023, doi: 10.31436/ijpcc.v9i1.285.
- [9] R. Singh and S. Singh, "Text Similarity Measures in News Articles by Vector Space Model Using NLP," *Journal of The Institution of Engineers (India): Series B*, vol. 102, no. 2, pp. 329–338, Apr. 2021, doi: 10.1007/s40031-020-00501-5.
- [10] S. H. Park and K. Kim, "Collaborative filtering recommendation system based on improved Jaccard similarity," *J Ambient Intell Humaniz Comput*, vol. 14, no. 8, pp. 11319–11336, Aug. 2023, doi: 10.1007/s12652-023-04647-0.
- [11] M. AlMousa, R. Benlamri, and R. Houry, "Exploiting non-taxonomic relations for measuring semantic similarity and relatedness in WordNet," *Knowl Based Syst*, vol. 212, p. 106565, Jan. 2021, doi: 10.1016/j.knosys.2020.106565.
- [12] D. Zhang, X. You, S. Liu, and K. Yang, "Multi-Colony Ant Colony Optimization Based on Generalized Jaccard Similarity Recommendation Strategy," *IEEE Access*, vol. 7, pp. 157303–157317, 2019, doi: 10.1109/ACCESS.2019.2949860.
- [13] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations," *Organ Res Methods*, vol. 25, no. 1, pp. 114–146, Jan. 2022, doi: 10.1177/1094428120971683.
- [14] T. Jang, J. Ahn, and S. L. Kim, "Punctuation restoration Model and Spacing Model for Korean Ancient Document," Dec. 2023, [Online]. Available: <http://arxiv.org/abs/2312.11881>
- [15] A. Irsyad and N. A. Rakhmawati, "Community detection in twitter based on tweets similarities in indonesian using cosine similarity and louvain algorithms," *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 6, no. 1, pp. 22–31, 2020, doi: 10.26594/register.v6i1.1595.
- [16] J. Praveen Gujjar, H. R. Prasanna Kumar, and M. S. Guru Prasad, "Advanced NLP Framework for Text Processing," in *2023 6th International Conference on Information Systems and Computer Networks (ISCON)*, IEEE, Mar. 2023, pp. 1–3. doi: 10.1109/ISCON57294.2023.10112058.