Calculation of Similarity between MUI Fatwas: A Comparison of Word Extraction and String-Matching Algorithms

Mohamad Fahmi Syafiudin[a]*, Gagatsatya Adiatmaja[a], Bilal Hidayaturrohman[a]

[a] Information System Department, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia.

ABSTRACT

Fatwas, as religious rulings issued by the Indonesian Ulama Council (MUI), play a crucial role in guiding the Muslim community. This research aims to analyze the similarity between these fatwas, contributing to the field by comparing various similarity methods. The dataset includes 380 fatwa titles collected from the official website of the National Sharia Council of the Indonesian Ulama Council. The research follows a structured methodology: starting with data collection, followed by text pre-processing involving punctuation removal, stemming, and stop word elimination. Word extraction techniques such as Bag of Words (BoW), TF-IDF (Term Frequency-Inverse Document Frequency), and BERT (Bidirectional Encoder Representations from Transformers) are then applied. Similarity is calculated using Jaccard Similarity, Cosine Similarity, Euclidean Distance, and Dice Coefficient. The results show that Cosine Similarity combined with TF-IDF achieves the highest performance with an F1 Score of 0.299. This study is novel in its comprehensive comparison of multiple similarity methods applied to MUI fatwas, providing valuable insights for researchers and practitioners in Natural Language Processing (NLP).

Keywords: *Fatwa MUI, Word Extraction,* Jaccard Similarity, Cosine Similarity, Euclidean Similarity, Dice Similarity.

© 2025 Pusat Kajian Halal ITS. All rights reserved.

1. Introduction

When a Muslim has a question about Islam, they can seek a fatwa from an Islamic scholar. Fatwa is a legal opinion based on Islamic teachings, provided by an expert in Islamic law. In Islam, there are four sources of law in formulating fatwas: the Quran, the Sunnah (actions, sayings, and approvals of Prophet Muhammad), consensus among scholars, and personal reasoning if there is no evidence from the previous three sources. All actions are considered valid unless there is a fatwa stating otherwise based on one of these four sources [1] Fatwas have a significant role in the daily lives of Muslims as they offer religious guidance on various aspects of life [2]. They help clarify religious obligations and provide solutions to

* Corresponding author.
Email address: fahmisyaifudin00@gmail.com

Received May 13th, 2024; Revised June 28th, 2024; Accepted February 6th, 2024

contemporary issues that may not be explicitly addressed in primary religious texts. The authority of a fatwa depends on the credibility and knowledge of the scholar issuing it, as well as the acceptance within the Muslim community [3].

The continuity of Islamic law in Indonesia heavily relies on fatwas issued by individuals and institutions. One of the institutions playing a crucial role in this regard is the Indonesian Ulama Council (MUI) [4]. The main task of the MUI is to issue fatwas, which are responses and answers to questions and issues within the Muslim community. The process of issuing fatwas at the MUI involves comprehensive studies and in-depth discussions among scholars in the Fatwa Commission. MUI fatwas are based on the Quran, Hadith, and the scholarly interpretation (ijtihad) of the ulama [5].

The similarity of MUI fatwas needs to be determined to identify whether there are fatwas that are actually duplicates or very similar to those that already exist, thus helping the relevant parties in managing and simplifying the fatwa archives. However, comparing MUI fatwas presents a formidable challenge. Previous research has explored diverse approaches to text similarity across different domains, highlighting significant parallels. For instance, Akhmad Irsyad et al.'s investigation into community detection on Twitter using cosine similarity and the Louvain algorithm underscores the complexities of textual analysis in dynamic contexts [6]. Additionally, Rakhmawati et al. employed methodologies like Levenshtein distance and WordNet-based conceptual similarity to address data redundancy in food ingredient databases, revealing gaps in research applicable to our study [7].

This research endeavors to assess various similarity methods like Jaccard Similarity, Cosine Similarity, Euclidean Distance, and Dice Coefficient for comparing MUI fatwas, aiming to identify the most effective approach. The study will evaluate each method's performance in analyzing fatwa similarity.

The approach used in this research is expected to provide a deeper understanding of the effectiveness of similarity methods in the case study of comparing MUI fatwas. By evaluating and comparing the performance of various methods, this research is expected to provide valuable insights for researchers and practitioners in selecting suitable similarity methods. This research aims to contribute new insights into measuring and comparing texts like MUI fatwas. This not only enhances the quality of analysis in related studies but also opens avenues for new applications in Natural Language Processing (NLP) and Islamic studies.

2. Related Works

Table 1. List of Research

Author	Word Extraction	Methods	Data
Rakhmawati et al. [7]	Bag-of-word	Jaccard Distance, Jaro-Winkler Distance, Levenshtein Distance, Wordnet	Food Products
Azmi Adi Firmansyah et al. [8]	TF-IEF	Euclidian Distance, Cosine Similarity	Product composition

(continued on next page)

Table 1. (continued)

Author	Word Extraction	Methods	Data
Akhmad Irsyad et al. [6]	Bag-Of-Words, word2vec	Cosine Similarity	Twitter
Ritika Singh [9]	Bag of Word, TF-IDF	Cosine, Jaccard, Euclidian Similarity,	Hindi & English News
Amr A. Munshi et al [10]	Bag-of-Words	LSTM, GRU	Islamic Fatwas
Reda Ahmed Zayed et [1]	Bag-of-Words	HOMER	Islamic Fatwas
Hasnawati et al [11]	SVM, IndoBERT	Cosine Similarity, POS Tag	Questions On Indonesian Language Subjects

Table 1 illustrates related research on text similarity [6], [7], [8], [9] utilizing diverse such as food product, twitter, and news datasets. However, for studies involving Islamic fatwas [1], [10] they primarily focus on classification and lack a comparison of similarities between fatwa titles. In this study, various text similarity methods and word extraction techniques from the previously mentioned related studies are implemented.

To determine the similarity method to be used, previous research has been conducted to find similarities in food ingredients across many products [12]. The results showed that the cosine similarity method is better than Euclidean distance. Previous research has also been conducted to quickly summarize documents suspected of having many similar words [13]. The results showed that Jaccard similarity is better than cosine similarity and dice coefficient. Therefore, it is necessary to verify the results obtained based on the methods used in previous studies, namely cosine similarity, Euclidean distance, Jaccard similarity, and dice coefficient with different datasets.

Then, to determine word extraction, previous research proved that TF-IDF similarity results are better than Bag of Words [13], and other research has shown that IndoBERT has much better results than SVM [11]. Therefore, it is necessary to evaluate and compare the word extraction methods previously used, namely TF-IDF, Bag of Words, and IndoBERT.

3. Method

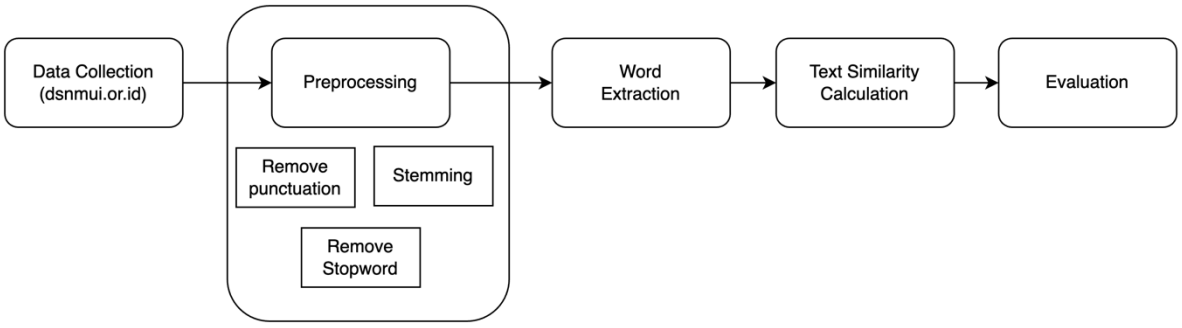

Fig. 1 Research Workflow diagram

Fig. 1 illustrates the system, where all MUI fatwas are compared to each other. If the similarity score between fatwas is higher than 70%, they are considered similar. Our method consists of four processes: data collection, pre-processing, word extraction, and similarity calculation.

3.1. Data Collection

Our data is obtained through the process of scraping halal fatwas from the official website of the National Sharia Council of the Indonesian Ulama Council (https://dsnmui.or.id/kategori/fatwa/). The collected data includes 380 fatwa titles, fatwa themes, fatwa numbers, and dates of issuance. A summary of the MUI fatwa data is presented in Table 2.

Table 2. Example of Raw dataset

Fatwa Title	Fatwa Theme	Fatwa Number	Date Set
Panduan Penyelenggaraan Ibadah Di Bulan Ramadan Dan Syawal 1442 H (Guidelines for Observing Worship in the Month of Ramadan and Syawal 1442 H)	Sosial Kemasyarakatan (Social and Community Affairs)	Keputusan Ijtima Ulama Komisi Fatwa Se-Indonesia Tahun 2003 (Decision of the National Ulama Ijtima of the Fatwa Commission of Indonesia in 2003)	16 December 2003
Penggunaan Mikroba Dan Produk Mikrobial Dalam Produk Pangan (Use of Microbes and Microbial Products in Food Products)	POM Iptek (Science and Technology Assessment Board)	No. 01 of 2010	19 January 2010
Penggunaan Vaksin Meningitis Bagi Jemaah Haji Atau Umrah (Use of Meningitis Vaccine for Hajj or Umrah Pilgrims)	POM Iptek (Science and Technology Assessment Board)	No. 06 of 2010	16 July 2010
Pensucian Alat Produksi Yang Terkena Najis Mutawassithah (Najis Sedang) Dengan Selain Air (Purification of Production Equipment Contaminated with Medium-Level Impurities (Mutawassithah) with Non-Water Substances)	POM Iptek (Science and Technology Assessment Board)	No. 09 of 2011	3 March 2011
Cara Pensucian Ekstrak Ragi Dari Sisa Pengolahan Bir (Method of Purifying Yeast Extract from Beer Processing Residue)	POM Iptek (Science and Technology Assessment Board)	No. 10 of 2011	4 March 2011

3.2. Pre-processing

The collected data is then extracted to obtain the fatwa titles for further preprocessing. Preprocessing is a series of steps used to prepare raw text data for effective processing in Natural Language Processing (NLP) [14]. Preprocessing in this study consists of three stages. The first stage involves removing punctuation such as commas, question marks, periods, exclamation marks, parentheses, and numbers [14]. The second stage is changing to the base form of words, which is done by removing prefixes, suffixes (stemming), and affixes from each word [14]. The third stage involves removing stop words, which are common words such as "and", "or", "in", "to", "that", and similar ones that frequently appear in the language and do not have significant meaning when analyzed. An example dataset after text preprocessing is shown in Fig. 2.

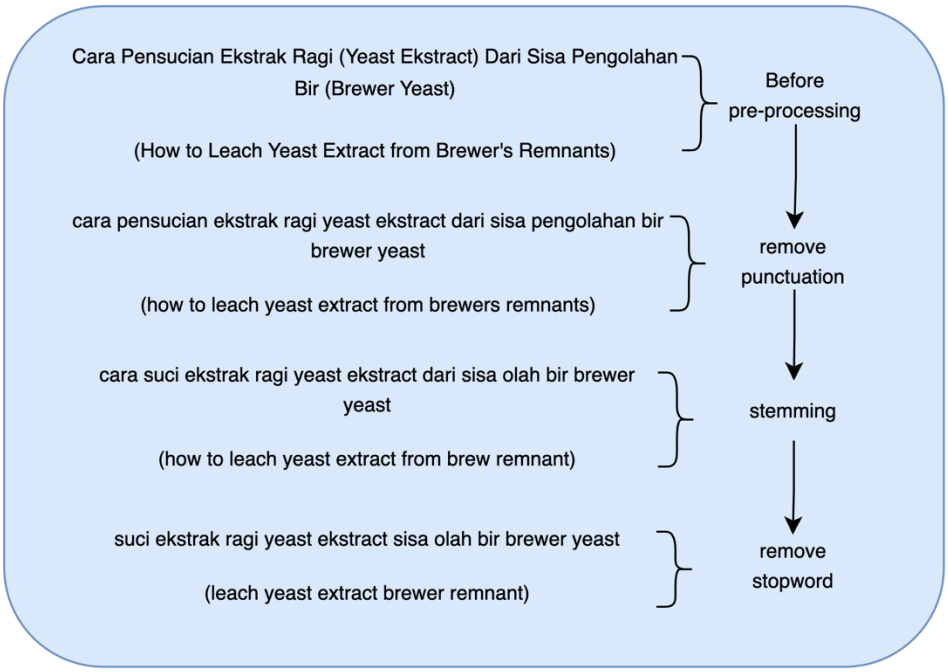

Fig. 2 Example data before preprocessing and after pre-processing

3.3. Word Extraction

Important information in textual data needs to be transformed into numerical vectors before it can be used in similarity calculation processes or machine learning models. Word extraction has several methods, some of which are used including Bag of Words, TF-IDF (Term Frequency and Inverse Document Frequency), and the pre-trained language model BERT.

a. Bag of Word

Bag of Word (BoW) is a method used for word extraction. This method is highly effective and has advanced capabilities in selecting and classifying features by creating a bag for each type of instance. The BoW model is described as a simplified representation frequently used in Natural Language Processing (NLP) and Information Retrieval (IR). This model depicts text such as sentences or documents as a collection of its words, only considering the presence of duplicate words and disregarding grammatical structure and word order. The BoW model is

widely used in document classification, where features for training classification are created based on the presence or frequency of each word [15].

b. BERT

In 2018, Google developed BERT (Bidirectional Encoder Representations from Transformers), a pre-trained language model. This model is built using complex neural network architectures and equipped with attention-focused components. Its purpose is to handle and understand sequential data, such as text, and to learn how words relate to each other in context [16].

IndoBERT is a model that utilizes transformer technology and is trained through the Huggingface platform with the standard BERT-Base configuration that does not differentiate between uppercase and lowercase letters. This model is developed using over 220 million words from three primary sources: Wikipedia with 74 million words, news stories from Kompas, Tempo, and Liputan6 totaling 55 million words, and from an Indonesian corpus website providing 90 million words [17].

c. TF-IDF

TF-IDF is a classical method for weighting terms in text retrieval and data analysis processes automatically. This method determines the importance of a word in text using two measures, namely TF and IDF. TF refers to the number of occurrences of a term in a document, while IDF is a measure that determines how rare the term appears across all documents [13]. The formula for this measurement is:

$$IDF = \log\left(\frac{TN}{1+df}\right) \quad (1)$$

Where TN represents the total number of documents, while df is the number of documents containing a particular term. The TF-IDF value is calculated by multiplying TF by IDF.

3.4. Text Similarity Calculation

Various approaches exist to measure similarity between texts. These methods are divided into four categories: string-based, corpus-based, knowledge-based, and hybrid [18].

In string-based methods, there are two main types: sequence-based and token-based similarity. The type used in this study is token-based. Token-based methods analyze text similarity by breaking it down into small units called tokens, such as words. Each token is assigned a numerical label, and texts are considered similar if they have the same or similar tokens [18]. The application of token-based methods in this study involves using bag-of-words and TF-IDF to extract features from the text, and employing text similarity algorithms such as Jaccard Similarity, Dice Coefficient, Cosine Similarity, and Euclidean Distance to measure similarity between texts.

a. Jaccard Similarity

Jaccard similarity measures the similarity between sets by comparing the size of their intersection to the size of their union [7]. The formula for Jaccard similarity is:

$$J(A,B) = \frac{A \cap B}{A \cup B} = \frac{A \cap B}{|A|+|B|-|A \cap B|} \quad (2)$$

b. Cosine Similarity

Cosine Similarity measures the cosine of the angle between two vectors in an n-dimensional space. This similarity is calculated by taking the dot product of those vectors and dividing it by the product of their magnitudes [9]. The formula for this measurement is:

$$C(A,B) = \frac{A.B}{||A|| \times ||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (3)$$

Cosine angle indicates the level of similarity. If the angle is 0 degrees, the cosine value is 1, indicating that the vectors are identical. As the angle deviates from 0, the cosine value decreases, indicating lower similarity. If the angle is 180 degrees, resulting in a cosine value of -1, the vectors are completely dissimilar.

c. Euclidian Similarity

Euclidean distance, also known as L2 distance or Euclidean norm, is a method used in vector space models to assess similarity not based on angle measurement, but through direct linear distance between vectors. To determine how close the two points are, one would use Euclidean distance. This measurement implies that the closer the points are to each other, the more similar they are [9]. The formula for this measurement is:

$$E(A,B) = \sqrt{(A_2 - A_1)^2 + (B_2 - B_1)^2} \quad (4)$$

d. Dice Coefficient

The Dice Coefficient is almost the same as Jaccard similarity, where the formula utilizes the intersection of the two sets of words. The formula for this measurement [19]:

$$D(A,B) = \frac{2|A \cap B|}{|A|+|B|} \quad (5)$$

3.5. Evaluation

In data mining, several important metrics for evaluating model performance are accuracy, precision, recall (sensitivity), and F1 score. These metrics use four basic terminologies: TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) [20].

- True Positive is the condition when the data initially labeled as "similar" is correctly predicted as "similar."
- True Negative is the condition when the data initially labeled as "not similar" is correctly predicted as "not similar."
- False Positive is the condition when the data initially labeled as "not similar" is incorrectly predicted as "similar."
- False Negative is the condition when the data initially labeled as "similar" is incorrectly predicted as "not similar."

Precision is a measure that evaluates the model's ability to produce the correct number of predictions for the positive class compared to all results predicted as positive [21]. The formula for Precision:

$$Precission = \frac{TP}{TP+FP} \quad (6)$$

Where:

TP is True Positive

FP is False Positive

Sensitivity, or Recall, is a measure that describes the model's ability to accurately identify all true positive cases from the entire positive data [21]. The formula for Recall:

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

Where:

TP is True Positive

FN is False Negative

The F1 Score is a metric that provides the harmonic mean value of precision and recall, reflecting a balance between these two metrics and providing an overview of the overall effectiveness of the model in classifying the positive class [21]. The formula for the F1 Score:

$$F1\ Score = \frac{2*(Precission*Recall)}{Precission+Recall} \quad (8)$$

4. Results and Discussion

4.1. Results

This research conducted experiments to compare the level of similarity between MUI Fatwas, as shown in Table 3 and Table 4, by applying three different word extraction approaches. First, using the token-based similarity method with Bag-of-Words and text similarity algorithms Jaccard Similarity and Dice Coefficient. Token-based similarity also employs TF-IDF and text similarity algorithms Cosine Similarity and Euclidean Distance. Second, using the knowledge-based method with IndoBert and text similarity algorithms Cosine Similarity and Euclidean Distance.

Table 3. List of sample fatwa title

Code	Fatwa
F008	Amil Zakat (Zakat Administrator)
F346	Pengelolaan Zakat (Zakat Management)
F015	Badal Thawaf Ifadhah (Pelaksanaan Thawaf Ifadhah Oleh Orang Lain) (Substitute Thawaf Ifadhah (Performing Thawaf Ifadhah on Behalf of Someone Else))
F241	Taswiyah Al-Manhaj (Penyamaan Pola Pikir Dalam Masalah-Masalah Keagamaan) (Harmonization of Methodologies (Unifying Perspectives in Religious Matters))
F024	Penggunaan Plasenta Hewan Halal Untuk Bahan Obat (Use of Halal Animal Placenta for Medicine Ingredients)

(continued on next page)

Table 3. (continued)

Code	Fatwa
F057	Penggunaan Alkohol / Etanol Untuk Bahan Obat (Use of Alcohol/Ethanol for Medicine Ingredients)
F061	Transplantasi Organ Dan Atau / Jaringan Tubuh Untuk Diri Sendiri (Organ and/or Tissue Transplantation for Oneself)
F062	Transplantasi Organ Dan/Atau Jaringan Tubuh Dari Pendonor Hidup Untuk Orang Lain (Organ and/or Tissue Transplantation from a Living Donor to Another Person)
F081	Pendaftaran Haji Saat Usia Dini (Registering for Hajj at a Young Age)
F315	Pendaftaran Haji Usia Dini (Early Age Hajj Registration)

Table 4. Summary of similarity results

Fatwa Code	Fatwa Code	IndoBERT		TF-IDF		Bag of Word		Labeling
		Cosine similarity	Euclidian distance	Cosine Similarity	Euclidian distance	Jacard Similarity	Dice Coefficient	
F008	F346	0.800	0.119	0.324	0.462	0.333	0.500	Not Similar
F015	F241	0.807	0.109	0.0	0.414	0.000	0.000	Not Similar
F024	F057	0.805	0.120	0.362	0.469	0.285	0.444	Not Similar
F061	F062	0.834	0.129	0.712	0.568	0.500	0.666	Similar
F081	F315	0.846	0.136	1.00	1.000	1.000	1.00	Similar

The evaluation of similarity between fatwas is conducted using several parameters, namely precision, recall, and F1-score [21]. These parameters are measured by comparing the results of text extraction methods and string-matching algorithms with manually labeled fatwa similarity.

Table 5. Method evaluation of text extraction and string-matching algorithms

Method	Precision	Recall	F1-Score
IndoBERT Cosine Simillarity	0.026	0.169	0.045
IndoBERT Euclidian Distance	0.000	0.000	0.000
TF-IDF Cosine Simillarity	0.956	0.177	0.299
TF-IDF Euclidian Distance	1.000	0.016	0.031
Bag of Word Jacard Simmilarity	1.000	0.048	0.092
Bag of Word Dice Coefficient	0.947	0.145	0.251

Based on Table 5, the Cosine Similarity method with TF-IDF has a higher F1-Score, which is 0.299, in measuring the similarity between fatwas compared to other text extraction methods and string-matching algorithms. Evaluating fatwa title similarity might result in a lower-than-expected F1 score due to class imbalance. This happens because there are many more

dissimilar titles (negative examples) compared to similar titles (positive examples). In machine learning, F1 score balances precision (finding correct similar titles) and recall (finding all similar titles). With class imbalance, true positives (correct similar titles) are naturally less frequent than true negatives (correct dissimilar titles). This skews the F1 score towards the majority class (dissimilar titles), leading to a lower overall score. In addition, TF-IDF, and BoW word extraction has high precision and low recall, it means the word extraction is good at identifying true positive results, but it misses at identifying false negatives results. On the other hand, IndoBERT produces low precision, recall, and F-1 scores, indicating that word extraction using IndoBERT has poor overall performance.

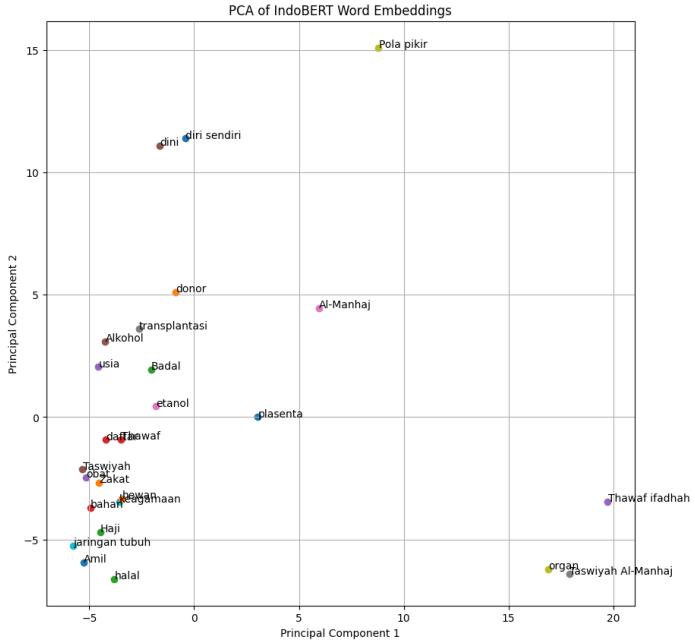

Fig. 3 IndoBERT word similarity

4.2 Discussion

The text extraction method using IndoBERT yields a low precision, recall, and F1 score. This occurs because this method disregards sentence context and focuses on literal word similarity. For example, from Fig. 3, words like "Thawaf Ifadhoh" and "Taswiyah Al-Manhaj" are considered similar because both are loanwords from Arabic. However, their meanings in sentence context can differ. Consequently, this approach often produces irrelevant or contextually inappropriate text extractions, unless fine-tuning the BERT model with a suitable dataset.

The Euclidean Distance algorithm shows suboptimal performance with low F1 scores across all text extraction methods. This is due to its main limitation, which is the inability to scale values. Unlike other string-matching algorithms with a scale of 0-1, Euclidean Distance lacks

a standardized scale. Therefore, proper normalization is required to enhance Euclidean Distance performance in the context of string matching.

5. Conclusion

This research aims to compare the similarity between MUI fatwas using four text similarity methods: Jaccard Similarity, Cosine Similarity, Euclidean Distance, and Dice Coefficient. The results indicate that Cosine Similarity with word extraction approach using TF-IDF has the highest performance for calculating the similarity of MUI fatwas. The word extraction method using IndoBERT yields low performance for MUI fatwa calculation, because IndoBERT focuses on literal word similarity with a limited dataset. In contrast, traditional extraction methods like TF-IDF and Bag-of-words are good at finding truly similar titles indicate by high precision, but they miss many others indicated by low recall. This happens because the data they use isn't balanced, so they can't catch all the similarities. Meanwhile, the Euclidean Distance string matching algorithm performs poorly due to its lack of a standardized scale and improper normalization.

In future work, the calculation of similarity between MUI fatwas will be refined to provide more accurate and insightful comparisons. A key enhancement will be the calculation of the F1 score with a balanced dataset. Additionally, we plan to fine-tune a BERT model using a corpus of data from Muslim news portals. This specialized training will enable the model to capture the nuances of religious language and context more effectively. Unlike previous approaches that focused solely on the titles of the fatwas, we suggest for analyze the entire content to identify deeper and more meaningful similarities. This comprehensive approach is expected to yield better insights in the study of MUI fatwas.

References

[1] R. A. Zayed, M. F. A. Hady, and H. Hefny, “Islamic fatwa request routing via hierarchical multi-label Arabic text categorization,” in *Proceedings - 1st International Conference on Arabic Computational Linguistics: Advances in Arabic Computational Linguistics, ACLing 2015*, Institute of Electrical and Electronics Engineers Inc., Feb. 2016, pp. 145–151. doi: 10.1109/ACLing.2015.28.

[2] J. A. Ali, “Contemporary Islamic Revivalism: Key Perspectives,” 2012.

[3] W. B. Hallaq, *A history of Islamic legal theories : an introduction to Sunnī uṣūl al-fiqh*. 1997.

[4] Nasrullah, “MAJELIS ULAMA INDONESIA (MUI); STUDI ATAS PENGGUNAAN METODOLOGI QIYAS SEBAGAI UPAYA PENETAPAN HUKUM ISLAM DI INDONESIA,” 2017.

[5] M. Asad, “Ulama in Indonesian Politics: Analysis on the Attitudes of The Majelis Ulama Indonesia (MUI) on the General Elections,” *Akademika*, vol. 16, no. 1, Jun. 2022, doi: 10.30736/adk.v16i1.764.

[6] 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.

[7] N. Aini Rakhmawati and M. Jannah, "Food ingredients similarity based on conceptual and textual similarity," 2021. [Online]. Available: http://halal.addi.is.its.ac.id/

[8] N. Aini Rakhmawati, A. Adi Firmansyah, P. Maulidya Effendi, R. Abdillah, and T. Agung Cahyono, "Auto Halal Detection Products Based on Euclidian Distance and Cosine Similarity," vol. 8, pp. 4–6, 2018, [Online]. Available: http://halal.addi.is.its.ac.id;

[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] A. A. Munshi, W. H. AlSabban, A. T. Farag, O. E. Rakha, A. A. Al Sallab, and M. Alotaibi, "Towards an Automated Islamic Fatwa System: Survey, Dataset and Benchmarks," *International Journal of Computer Science and Mobile Computing*, vol. 10, no. 4, pp. 118–131, Apr. 2021, doi: 10.47760/ijcsmc.2021.v10i04.017.

[11] Hasmawati and Ade Romadhony, "Similar Questions Identification on Indonesian Language Subject Using Machine Learning," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 12, no. 2, pp. 196–202, Jul. 2023, doi: 10.23887/janapati.v12i2.62582.

[12] N. Aini Rakhmawati, A. Adi Firmansyah, P. Maulidya Effendi, R. Abdillah, and T. Agung Cahyono, "Auto Halal Detection Products Based on Euclidian Distance and Cosine Similarity," vol. 8, pp. 4–6, 2018, [Online]. Available: http://halal.addi.is.its.ac.id;

[13] 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.

[14] M. J. Sulastri, N. Aini Rakhmawati, and R. Indraswari, "Identifying Gender Bias in Online Crime News Indonesia Using Word Embedding," in *2023 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation, ICAMIMIA 2023 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 774–778. doi: 10.1109/ICAMIMIA60881.2023.10427911.

[15] Wisam A. Qader, Musa M.Ameen, and Bilal I. Ahmed, "An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges," in *Fifth International Engineering Conference on Developments in Civil & Computer Engineering Applications 2019 - (IEC2019) - Erbil - IRAQ*, 2019.

[16] K. A. Alshaikh, O. A. Almatrafi, and Y. B. Abushark, "BERT-Based Model for Aspect-Based Sentiment Analysis for Analyzing Arabic Open-Ended Survey Responses: A Case Study," *IEEE Access*, vol. 12, pp. 2288–2302, 2024, doi: 10.1109/ACCESS.2023.3348342.

[17] A. B. Y. A. Putra, Y. Sibaroni, and A. F. Ihsan, "Disinformation Detection on 2024 Indonesia Presidential Election using IndoBERT," in *2023 International Conference on*

Data Science and Its Applications, ICoDSA 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 350–355. doi: 10.1109/ICoDSA58501.2023.10277572.

[18] D. D. Prasetya, A. P. Wibawa, and T. Hirashima, "The performance of text similarity algorithms," *International Journal of Advances in Intelligent Informatics*, vol. 4, no. 1, pp. 63–69, Mar. 2018, doi: 10.26555/ijain.v4i1.152.

[19] S. P. Pati and R. Rautray, "An Empirical Analysis of Similarity based Single Document Summarization," in *Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021*, Institute of Electrical and Electronics Engineers Inc., Apr. 2021, pp. 860–864. doi: 10.1109/ICCMC51019.2021.9418297.

[20] S. A. Khan and Z. Ali Rana, "Evaluating Performance of Software Defect Prediction Models Using Area Under Precision-Recall Curve (AUC-PR)," in *2019 2nd International Conference on Advancements in Computational Sciences (ICACS)*, IEEE, Feb. 2019, pp. 1–6. doi: 10.23919/ICACS.2019.8689135.

[21] C. Anwar ul Hassan, M. Sufyan Khan, and M. Ali Shah, "Comparison of Machine Learning Algorithms in Data classification," in *Proceedings of the 24th International Conference on Automation & Computing*, 2018.