Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 25, No. 1, November 2025, pp. 63∼72
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 10/C/C3/DT.05.00/2025
DOI: 10.30812/matrik.v25i1.5600

Comparative Analysis of TF-IDF and Modern Text Embedding for the Classification of Islamic Ideologies on Indonesian Twitter

Siti Ummi Masruroh[1], **Cong Dai Nguyen**[2], **Doni Febrianus**[1]
[1]Universitas Islam Negeri Syarif Hidayatullah, Jakarta, Indonesia
[2]Le Quy Don Technical University, Hanoi 10000, Vietnam

Article Info

Article history:

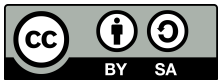
Received August 26, 2025
Revised October 02, 2025
Accepted October 08, 2025

Keywords:

Islamic Ideologies;
Machine Learning;
Social Media;
Support Vector Machine;
Text Classification.

ABSTRACT

The ideological polarization that has emerged on social media platforms like Twitter, particularly regarding discussions on Islamic ideologies in Indonesia, has led to the rapid spread of da'wah. However, it has also created challenges in effectively classifying tweets into distinct Islamic ideologies, such as Liberal Islam and Moderate Islam (Wasathiyyah). The lack of effective methods for accurately classifying such nuanced content presents a significant challenge. To address this problem, the research aimed to develop and evaluate a machine learning model that compares the effectiveness of traditional word vectorization methods (TF-IDF) with modern text embedding models (Nomic Embed v2). The study utilized the Knowledge Discovery in Databases (KDD) framework, scraped relevant data using the Twitter API, and annotated the dataset based on ideology. Preprocessing techniques such as case folding, stopword removal, and symbol removal were applied to the dataset. Classification was carried out using an SVM model, and cross-validation was employed to assess the model's accuracy. The findings indicate that the embedding model improved the accuracy by providing nuanced semantic context for the tweets, suggesting that modern semantic models can outperform traditional methods in classifying complex, context-dependent texts.



Corresponding Author:

Siti Ummi Masruroh,
Faculty of Science and Technology and Study Program Informatics Engineering,
Universitas Islam Negeri Syarif Hidayatullah, Jakarta, Indonesia,
Email: ummi.masruroh@uinjkt.ac.id.

How to Cite:
S. U. Masruroh, Cong D. Nguyen, and D. Febrianus, "Comparative Analysis of TF-IDF and Modern Text Embedding for the Classification of Islamic Ideologies on Indonesian Twitter", *MATRIK: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, Vol. 25, No. 1, pp. 63-72, November, 2025.

1. INTRODUCTION

In the last two decades, the landscape of religious communication in Indonesia has undergone a significant transformation. Fast technological improvements have shifted the discussion space from the traditional spaces, such as mosques, to a more open and dynamic space, such as the digital medium. Social media has become the primary platform for discussion, community, and a stage for dialogue and debates [1]. Of all the platforms, Twitter has been the most popular room for discussion, mainly due to the real-time and anonymous characteristics of the platform, allowing for a more open discussion, and features such as hashtags facilitate the amplifying of certain topics quickly [2].

This phenomenon has given rise to digital da'wa, where religious influencers and Islamic organizations utilize these platforms to spread their teachings and build a following on a scale much larger than that in physical spaces [3]. Now individuals have direct access to various religious interpretations without interference from traditional institutions. However, the freedom granted from the digital space has spawned a paradox. On the one hand, social media has successfully expanded the reach of religious influencers to a global stage. On the other hand, the algorithmic architecture that social media platforms use has been the main propeller for ideological fragmentation and social polarization, specifically between Liberal Islam and Moderate Islam (Wasatiyah).

Liberal Islamic thought in Indonesia, which gained significant momentum in the post-New Order era, can be identified by several distinctive characteristics. One of its main pillars is the idea of secularism, which advocates a separation between worldly (political) and otherworldly (religious) authority [4]. Adherents of this school of thought reject the idea of a formal Islamic state and believe that the form of the state is a product of human ijtihad, not a divine mandate.

The second pillar is religious pluralism, which involves not only tolerating the existence of other religions but also adopting a theological perspective that views all religions as valid paths to God, thus considering truth to be relative. When interpreting sacred texts, liberal thought tends to employ a hermeneutical approach, prioritizing the "religious-ethical spirit" of a text over its literal meaning. Furthermore, this movement consistently advocates for progressive issues such as gender equality, defending the rights of minority groups, and freedom of opinion and expression [5, 6]. This movement is often associated with organizations such as the Liberal Islam Network (JIL) and intellectual figures such as Ulil Abshar Abdalla.

Moderate Islam, or Wasathiyyah, positions itself as a distinctive middle path, not as a compromise between liberalism and extremism, but as an essential manifestation of Islamic teachings themselves [7]. This concept is rooted in the principle of Tawassuth, which is the attitude of taking a middle path and avoiding extremism (ghuluw), both towards excessive liberalism and rigid radicalism or fundamentalism. Another fundamental principle is Tawazun (balance), which encompasses the balance between worldly affairs and the afterlife, between reason and revelation, and between sacred texts and the context of social reality. Other key characteristics include Tasamuh (tolerance of differences), I'tidal (being upright, just, and objective), and Musawah (equality), all of which aim to realize Islam as rahmatan lil 'alamin (blessing for all nature). This ideology is embraced by Indonesia's largest Islamic mass organizations, such as Nahdlatul Ulama (NU) and Muhammadiyah, and is often expressed through concepts such as "Islam Nusantara."

Although these two schools of thought appear fundamentally different, a significant computational challenge arises from the semantic overlap. Terms such as "tolerance," "pluralism," "justice," and "humanity" are frequently used by both groups. The crucial difference lies not in the words themselves, but in the conceptual and contextual frameworks surrounding them. Liberal Islam tends to frame these terms within the discourse of universal human rights, rationalism, and individual freedom. In contrast, Moderate Islam places these same concepts within a specific Islamic theological framework, such as the maqasid al-shari'ah (objectives of sharia) and the principles of Wasatiyyah.

Considering the large volume of discussion and its effect on social cohesion, classifying these two beliefs can be challenging if done manually. In the Indonesian context, M. Mudhofi et al. [8] conducted a pioneering study, the first to use a data mining approach to analyze public opinion on moderate religion in Indonesia. Their research elaborated on three poles of religious understanding in Indonesia: fundamentalist, liberal, and moderate, each of which uses social media as a means to promote its teachings. By using text mining and sentiment analysis methods on Twitter data collected with keywords like "tolerance," "radical," and "diversity," they found that public sentiment toward the values of religious moderation was predominantly positive. This research provides a strong contextual foundation, demonstrating that analyzing religious discourse on Indonesian Twitter is a relevant field and amenable to computational analysis. In line with this, Nuwairah and Munsyi [9]. specifically classified Islamic preaching content on Indonesian websites to detect radical ideologies. They defined radical content in the Indonesian context as content that incites violence, spreads hatred (SARA), and opposes nationalism. Using the k-Nearest Neighbor (kNN) method on data from websites blocked by the Ministry of Communication and Informatics, they achieved an accuracy of 66.37%.

K. T. Mursi et al. [10]. also successfully developed a model to detect Islamic radicalism in Arabic tweets. Facing the challenge of a lack of public datasets, they compiled a dataset manually labeled by experts, consisting of 3,000 tweets. By applying careful data preprocessing and using the Support Vector Machine (SVM) algorithm with TF-IDF features, their model achieved a very high

accuracy of 92%. This success provides strong methodological justification for the use of SVMs in the nuanced task of classifying religious texts.

Using Naive Bayes and TF-IDF, Moustafa and Olowolayemo [11]. applied text classification methods to distinguish Muslim ideologies on websites, specifically classifying 60 websites into Sunni and Shia categories. They achieved 89% accuracy and successfully extracted the keywords that best differentiated the two groups.

Similarly, W. González-Baquero et al. [12]. also applied a large-scale computational approach to analyze conversations about Islam in Spain. They used topic modeling and sentiment analysis on 190,320 tweets to map the dominant topics and sentiments. The analysis showed that while negative topics contained Islamophobia, the majority of conversations were neutral and informative.

This research is important and fills a significant gap by addressing an issue in the current literature on religious discourse analysis, particularly in the context of Indonesia. While existing methods, such as TF-IDF, have been widely used for text classification, they primarily focus on statistical associations and keyword-based feature extraction, which may fall short in capturing the semantic and contextual nuances inherent in religious ideologies, especially when categorizing moderate versus liberal Islamic thoughts. The increasing use of social media has generated vast amounts of short, noisy texts, presenting a unique challenge for traditional methods.

By employing the Support Vector Machine (SVM) model, a proven algorithm for text classification, this research explores a more sophisticated approach to handling these complexities. The primary advantages of Support Vector Machines (SVMs) for text classification include being one of the most commonly used algorithms in text classification tasks, offering stable performance, providing excellent performance in high-dimensional spaces, effectively handling non-linearly separable data using the ”kernel trick” [13].

A core component of this research is a comparative analysis of the traditional TF-IDF method against a more advanced feature extraction strategy using Nomic v2, a word embedding model trained on a large corpus including Indonesian text, to determine the impact of semantic context on classification accuracy. Ultimately, this research seeks not only to refine the classification of Indonesian religious ideologies but also to assess whether modern, semantic methods provide a significant advantage in analyzing nuanced religious discourse over traditional, keyword-based approaches.

2. RESEARCH METHOD

Knowledge Discovery in Databases framework (KDD) was used in this research; KDD itself is a proper methodology to analyze and understand such huge amounts of data [14]. The process of the KDD framework is as follows:

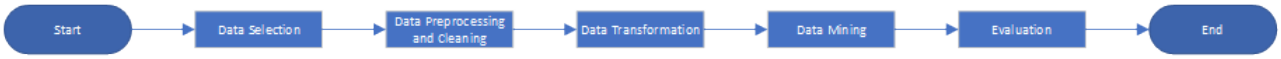


Figure 1. Knowledge Discovery in Databases Framework

Figure 1 outlines a typical KDD data mining process, beginning with data selection, where relevant data is identified for analysis. Next, data preprocessing and cleaning are performed to address issues such as missing values, duplicates, and noise, ensuring the data is in a proper format. After that, data transformation occurs, converting the data into a structure suitable for mining. In the data mining step, algorithms and techniques are applied to extract meaningful patterns and insights from the data. Following this, the evaluation stage assesses the quality of the mined data and the usefulness of the patterns discovered. Finally, the process concludes with the end of the workflow, signifying the completion of the data mining task.

2.1. Data Selection

The main dataset was acquired from Twitter using the tweetpy API. To ensure the data gathered from the API was relevant, a few criteria were set. This includes limiting the keyword to the following: Islam, Muslim, Al-Qur'an, Jihad, Radical, Liberal. Results were also filtered to include only tweets from users in Indonesia, based on coordinates, and the timeframe was set from 1 January 2011 to 1 January 2020.

2.2. Data Preprocessing and Cleaning

To reduce noise in the dataset, several preprocessing methods were employed. The preprocessing step is as follows: Tokenization, Filtering mentions(@username), URLs, Hashtags, Filtering non-alphanumeric symbols, Removing duplicate entries, Dropping rows with null values, Removing Indonesian stopwords (such as 'dari', 'ke', 'yang', 'di', etc.), Case folding. Stemming was not included in the preprocessing steps, as it is known that stemming could induce a loss of information and word variation, resulting in lower accuracy with the text classification model [15].

2.3. Data Transformation

Manual labeling was done on the preprocessed dataset. The first step is to do a literature analysis to derive the characteristics of each school of thought. The analysis draws on various academic literature, such as Dewi, Bustamam-Ahmad, and Agustina [16–18] that discuss Islamic thought in Indonesia, particularly those that differentiate between liberal, moderate, and other fundamentalist Islamic groups. Based on the results of the theoretical study, a structured annotation framework or guidelines was developed. This framework serves as an objective guide for annotators to ensure consistency in labeling. The characteristics for each category are defined as follows:

Table 1. Comparative Framework of Liberal and Moderate Islamic Thought in Indonesia

Dimension	Liberal Islam	Moderate Islam
The relation between religion and nationality	Supports secularism and divides authority between religion and politics	Supports Pancasila and rejects a formal Islamic state, while also accepting Islamic values in national life
Religious Pluralism	All religions are considered valid, and truth is relative	Tolerant of the existence of other religions but believes in the truths of Islam
Gender Issues	Supports gender equality fully	Progressive, supports women's role in society while adhering to the Sharia framework
Al-Qur'an Interpretation	Using a hermeneutic approach, emphasizing the ethical spirit over the literal text	Adhere to the Al-Qur'an and its hadith with a balanced interpretation between text and context

Data labeling was performed manually by two expert annotators, both lecturers at an Islamic university in Indonesia specializing in Islamic studies and communication. Each annotator independently reviewed the cleaned tweets. Based on the guidelines of the annotation framework, they assigned each tweet a label ("Liberal Islam" or "Moderate Islam"). This process also involves examining the context of the sentence and considering the potential for implicit meanings to avoid misinterpretation. Every tweet was given a class of 1 for tweets potentially containing Liberal Islam values, and 2 for Moderate Islam values.

The distribution of these two classes was also checked for oversampling or undersampling to remove potential biases. The labelled text was then converted to a vector using two methods for comparison: Term Frequency Inverse Document Frequency (TF-IDF). TF-IDF, as the name implies, works by calculating the relative frequency of words in a document and comparing it to the inverse proportion of that word over the entire document [19]. Words that have a large TF-IDF number tend to have a strong relationship with the document they appear in, implying that the word could be of interest in proportion to the whole document. Text Embedding: Text embeddings are a technique that encodes semantic information about sentence vectors, and thus could be used for data visualization, classification, and information retrieval. This technique excels in capturing meanings in text, where similar words could be positioned close to each other in a vector space. The text embedding model in question (nomic-embed-v2) uses existing bidirectional encoder representations from transformers (BERT) models as a base for producing text embeddings [20]. The models based on BERT can effectively read a series of words in either direction of the input text, and since it uses the attention mechanism to assign a word, its vector depends on the surrounding words [21]. The Nomic v2 model was specifically chosen because it was trained with 36.470.784 pairs of Indonesian text, making it suitable for capturing context in Indonesian datasets [20]. Its capability of multilingual information retrieval was also proven with benchmarks such as Miracl, which involves 1.446.315 Indonesian text passages [22]. The model scores 65.8 on the Miracl benchmark and outperforms other state-of-the-art BERT-based embedding models, such as Arctic Embed v2 base and mGTE Base. It also matches the performance of larger parameter models like Arctic Embed v2 Large, making it much more efficient in terms of computing.

2.4. Data Mining

From the labelled text, we applied a text classification model to determine the predicted class. The Support Vector Machine algorithm was chosen as the model for its high accuracy in text classification tasks. The SVM models were configured using the linear, radial, and polynomial kernels for comparison.

2.5. Evaluation

The performance of the trained model was then evaluated using K-Fold-Cross-Validation. The labelled dataset is partitioned into 10 subsets of equal size(folds). In every iteration, nine folds were used to train the model, and one fold was used as a test dataset. The results will be displayed as a confusion matrix. Metrics such as accuracy, precision, and recall could also be derived from the confusion matrix

3. RESULT AND ANALYSIS

3.1. Data Selection

The process successfully queried 500.000 rows of raw tweets, with ten features that include tweet ID, username, number of retweets, number of likes, the tweet itself, user coordinates, and date created. From the raw dataset, we selected 10.000 rows of data that are relevant to the religious topics. The other features besides the tweet itself were also discarded.

3.2. Data Preprocessing

After removing duplicates and null entries in the dataset, a total of 9,812 rows remain and are used for further analysis. These cleaned data are then carried forward to the subsequent pre-processing stages. This step ensures that the dataset is more consistent and reliable for building the machine learning model.

Table 2. Filtering mentions(@username), URLs, and Hashtags Results

Before	After
@hramad @Ih4nd4y4ni Sola Scriptura ini bisa tidak sih diartikan ”kembali ke Quran dan hadits”?	Sola Scriptura ini bisa tidak sih diartikan ”kembali ke Quran dan hadits”?
@sahaLAS: Mungkin seperti yang di bahas di artikel ini . . . ”Mengkritisi slogan kembali ke #Quran dan #Sunnah” http://www.rumahfiqih.com/fikrah-123-mengkritisi-slogan-kembali-ke-al-quran-dan-sunnah.html	Mungkin seperti yang di bahas di artikel ini . . .” Mengkritisi slogan kembali ke Quran dan Sunnah”
Apa makna kembali ke Quran dan Hadist? fb.me/Sig67TVe	Apa makna kembali ke Quran dan Hadist?

Table 3. Filtering non-alphanumeric symbols Results

Before	After
@hramad @Ih4nd4y4ni Sola Scriptura ini bisa tidak sih diartikan ”kembali ke Quran dan hadits”?	Sola Scriptura ini bisa tidak sih di artikan kembali ke Quran dan hadits
@sahaLAS: Mungkin seperti yang dibahas di artikel ini . . . ”Mengkritisi slogan kembali ke #Quran dan #Sunnah” http://www.rumahfiqih.com/fikrah-123-mengkritisi-slogan-kembali-ke-al-quran-dan-sunnah.html	Mungkin seperti yang dibahas di artikel ini Mengkritisi slogan kembali ke Quran dan Sunnah
Apa makna kembali ke Quran dan Hadist? fb.me/Sig67TVe	Apa sih makna kembali ke Quran dan Hadist

Table 4. Removing Stopwords Results

Before	After
@hramad @Ih4nd4y4ni Sola Scriptura ini bisa tidak sih diartikan ”kembali ke Quran dan hadits”?	Sola Scriptura diartikan Quran hadits
@sahaLAS: Mungkin seperti yang dibahas di artikel ini . . . ”Mengkritisi slogan kembali ke #Quran dan #Sunnah” http://www.rumahfiqih.com/fikrah-123-mengkritisi-slogan-kembali-ke-al-quran-dan-sunnah.html	Dibahas artikel Mengkritisi slogan Quran Sunnah
Apa makna kembali ke Quran dan Hadist? fb.me/Sig67TVe	makna Quran Hadist

Table 5. Case Folding Results

Before	After
@hramad @Ih4nd4y4ni Sola Scriptura ini bisa tidak sih diartikan ”kembali ke Quran dan hadits”?	sola scriptura diartikan quran hadits
@sahaLAS: Mungkin seperti yang dibahas di artikel ini . . . ”Mengkritisi slogan kembali ke #Quran dan #Sunnah” http://www.rumahfiqih.com/fikrah-123-mengkritisi-slogan-kembali-ke-al-quran-dan-sunnah.html	dibahas artikel mengkritisi slogan quran sunnah
Apa makna kembali ke Quran dan Hadist? fb.me/Sig67TVe	makna quran hadist

3.3. Data Transformation

From the cleaned dataset, a total of 5088 tweets were labeled as Liberal Islam and 4724 tweets were labeled as Moderate Islam. Downsampling was done to the Liberal Islam class to balance the classes and prevent any biases, resulting in the final dataset containing 9448 rows with 4724 rows from each class. After each method of vectorization is done, the TF-IDF process creates a total of 22.000 features containing every word from the cleaned dataset. While the Text Embedding process created 768 features.

3.4. Data Mining

Training was done using the SVM Model with two different extraction features: TF-IDF and Text embedding using Nomic V2. The performance of every scenario mentioned above was evaluated based on the accuracy. The result is as follows:

Table 6. Accuracy Results

Model	Accuracy
SVM (Polynomial Kernel + Nomic v2 Embeddings)	80.72%
SVM (Linear Kernel + Nomic v2 Embedding)	80.45%
SVM (Linear Kernel + TF-IDF)	77.07%
SVM (Polynomial Kernel + TF-IDF)	70.59%
SVM (Radial Kernel + Nomic v2 Embeddings)	53.67%
SVM (Radial Kernel + TF-IDF)	52.79%

As shown in Table 6, there is a substantial variance in performance across the different configurations. The model employing a Polynomial kernel with Nomic v2 embeddings achieved the highest classification accuracy at 80.72%. The Linear kernel paired with the same embeddings yielded a nearly identical accuracy of 80.45%. In contrast, models utilizing the Radial Basis Function (RBF) kernel performed poorly, with accuracies only slightly above the 50% baseline for a binary classification task.

3.5. Evaluation

To understand the reasoning behind these results, it's important first to clarify what precision and recall mean in this context. Precision measures the proportion of true positive predictions among all instances where the model predicted a certain class. In contrast, recall measures the proportion of true positives correctly identified by the model among all actual instances of that class. A model with high precision may identify fewer relevant instances but is more accurate when it does. On the other hand, a model with high recall identifies most of the relevant instances but may incorrectly label more non-relevant instances as relevant.

Table 7. Confusion Matrix and Metrics for SVM with TF-IDF (Accuracy: 77.97% $\pm$ 1.67%)

Model	TP	FP	FN	TN	Precision (Class 1)	Recall (Class 1)	Precision (Class 2)	Recall (Class 2)
SVM (Linear + Nomic v2)	3637	794	1087	3979	82.04%	76.99%	78.54%	83.39%
SVM (Linear + TF-IDF)	3865	1327	759	3428	74.92%	83.59%	81.84%	72.05%
SVM (Polynomial + Nomic v2)	3705	886	939	3857	81.03%	80.12%	80.42%	81.32%
SVM (Polynomial + TF-IDF)	3344	1405	1300	3342	70.41%	70.79%	70.78%	70.40%
SVM (Radial + Nomic v2)	4706	4370	18	377	51.85%	99.62%	95.44%	7.94%
SVM (Radial + TF-IDF)	4691	4436	33	307	51.40%	99.30%	90.29%	6.47%

Further analysis, detailed in Table 7, reveals the nature of each model's predictive behavior. The top-performing Polynomial kernel model demonstrates balanced precision and recall for both Class 1 (Islamic Liberal) and Class 2 (Islamic Moderate), indicating no significant classification bias. Conversely, the RBF kernel models exhibit a critical failure while achieving over 99% recall for Class 1; their recall for Class 2 is exceptionally low ($< 8\%$).

A consistent trend was the superiority of Nomic v2 embeddings over TF-IDF. For both the Linear and Polynomial kernels, the use of embeddings resulted in a significant accuracy improvement, with the most substantial gain of over 10 percentage points observed for the Polynomial kernel.

The experimental results offer several key insights. The superior performance of the Nomic v2 embeddings over TF-IDF aligns with established findings in natural language processing [23]. Unlike the frequency-based, bag-of-words approach of TF-IDF, semantic embeddings capture the contextual and relational meaning of words, creating a much richer feature space for the classifier. The ability of the SVM, particularly with a Polynomial kernel, to leverage dense, semantic features to define a more effective non-linear decision boundary is the most likely explanation for the observed performance gains.

The strong performance of both the Polynomial and Linear kernels when paired with embeddings suggests that the classes are largely separable in the high-dimensional space created by the embeddings. The marginal difference in accuracy between them (0.27%) implies that a linear boundary is nearly sufficient, though a non-linear one provides a slight advantage. The computational efficiency of the Linear kernel may, therefore, make it a preferable choice in resource-constrained environments.

The failure of the RBF kernel models is a significant finding. RBF kernels are inherently powerful but are highly sensitive to hyperparameter settings (e.g., C and y). The observed behavior of strong bias towards one class is symptomatic of the model failing to learn a generalizable decision function for the given data distribution. It underscores the necessity of rigorous hyperparameter optimization when employing such complex kernels, without which the model can produce poor or skewed results.

4. CONCLUSION

This research was conducted to analyze the polarization of Islamic religious discourse on Twitter, particularly between liberal and moderate Islam in Indonesia. The primary objective was to develop and evaluate an automatic classification model capable of accurately distinguishing between these two schools of thought based on the textual content of tweets. Using the Knowledge Discovery in Databases (KDD) framework, relevant Indonesian-language tweet data was collected, cleaned, and manually labeled. Next, a Support Vector Machine (SVM) model was applied using two different vectorization approaches for comparison: Term Frequency-Inverse Document Frequency (TF-IDF), which is based on word frequency, and Nomic Embed v2, which is based on semantic representation.

The findings demonstrate that semantic-based feature representation (embedding) gave a measurable improvement, as compared to word frequency-based representation (TF-IDF) for ideological text classification tasks. For religious communication and Islamic studies, this work provides a framework to identify the linguistic patterns that characterize different streams of digital da'wah, offering insights into how religious ideas are articulated and contested in the modern public sphere. Discourse between liberal and moderate Islam is often indistinguishable not only from keywords but also from context, sentiment, and the relationships between words within a sentence. Nomic Embed v2's ability to capture these nuances allows SVM to form more accurate decision boundaries. Based on these findings, several future research directions can be explored, such as expanding the dataset with a more recent time span, involving more annotators with strict guidelines to improve label objectivity, and testing more sophisticated and fine-tuned social media model architectures, such as Transformer-based models (e.g., IndoBERT), specifically designed to understand the Indonesian language context.

5. ACKNOWLEDGEMENTS

Thanks to both the Faculty of Science and Technology and the Faculty of Islamic Studies and Communication for the assistance and resources provided during the completion of this manuscript.

6. DECLARATIONS

AI USAGE STATEMENT

The authors acknowledge that Artificial Intelligence tools, including ChatGPT developed by OpenAI, were utilized to support language refinement, grammar correction, and paraphrasing in the manuscript preparation process. The authors confirm that all ideas, data interpretations, and conclusions are their own and not generated by the AI tool.

AUTHOR CONTIBUTION

During the preparation of this work, the author(s) used Gemini 2.5 Pro for collecting references and summarizing experimental results. After using this tool/service, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content

FUNDING STATEMENT

S.U.M was responsible for writing, machine learning computation, experiment, manuscript revision, validation, and interpretation of data. The author approved the final version of this manuscript.

COMPETING INTEREST

PUSLITPEN UIN Jakarta funds this research.

COMPETING INTEREST

The authors declare that there are no conflicts of interest.

REFERENCES

[1] M. Murniati, "Ruang Publik dan Wacana Agama: Dinamika Dakwah di Tengah Polarisasi Sosial," vol. 1, no. 1, pp. 26–33, June,2025, https://doi.org/10.70742/khazanah.v1i1.260.

[2] S. R. I. Rezeki, Y. Restiviani, and R. Zahara, "Penggunaan Sosial Media Twitter dalam Komunikasi Organisasi (Studi Kasus Pemerintah Provinsi DKI Jakarta dalam Penanganan Covid-19)," vol. 4, no. 2, pp. 63–78, 2020, https://doi.org/10.18592/jils.v4i2.3812.

[3] S. Hudaa, N. Nuryani, and B. Sumadyo, "Pesan Dakwah Hijrah Influencer untuk Kalangan Muda di Media Sosial," vol. 17, no. 2, pp. 105–121, January,2023, https://doi.org/10.47651/mrf.v17i2.198.

[4] A. S. Amin and M. S. Syarifah, "Liberal Islam and Its Influences on the Development of Quranic Exegesis in Indonesia and Malaysia," vol. 22, no. 1, pp. 137–160, January,2021, https://doi.org/10.14421/qh.2021.2201-07.

[5] N. Rubani, "Elemen Islam Liberal dalam Idea Pembaharuan Islam Ahmad Wahib: Elements of Liberal Islam In Ahmad Wahib's Idea Of Islamic Reform," vol. 16, no. 1, pp. 9–21, May,2023, https://doi.org/10.53840/jpi.v16i1.235.

[6] A. Maksum, I. Abdullah, S. Mas'udah, and M. Saud, "Islamic Movements in Indonesia: A Critical Study of Hizbut Tahrir Indonesia and Jaringan Islam Liberal," vol. 17, no. 2, pp. 71–82, December,2022, https://doi.org/10.22452/JAT.vol17no2.6.

[7] A. Halim, H. Hosaini, A. Zukin, and R. Mahtum, "Paradigma Islam Moderat di Indonesia dalam Membentuk Perdamaian Dunia," vol. 1, no. 4, pp. 705–708, October,2022, https://doi.org/10.59004/jisma.v1i4.239.

[8] M. Mudhofi, I. Supena, A. Karim, S. Safrodin, and S. Solahuddin, "Public opinion analysis for moderate religious: Social media data mining approach," vol. 43, no. 1, pp. 1–27, May,2023, https://doi.org/10.21580/jid.v43.1.16101.

[9] N. Nuwairah and M. Munsyi, "Classification Content in Indonesian Website Da'wah using Text Mining for Detecting Islamic Radical Understanding:," February,2022, pp. 11–16, https://doi.org/10.2991/assehr.k.220206.002.

[10] K. T. Mursi, M. D. Alahmadi, F. S. Alsubaei, and A. S. Alghamdi, "Detecting Islamic Radicalism Arabic Tweets Using Natural Language Processing," vol. 10, pp. 72 526–72 534, July, 2022, https://doi.org/10.1109/ACCESS.2022.3188688.

[11] A. Olowolayemo and S. Moustafa Sharey Moustafa, "Classifying Muslim Ideologies from Islamic Websites using Text Analysis Based on Naive Bayes and TF-IDF," vol. 10, no. 1, pp. 8–15, January,2024, https://doi.org/10.31436/ijpcc.v10i1.321.

[12] W. González-Baquero, J. J. Amores, and C. Arcila-Calderón, "The Conversation around Islam on Twitter: Topic Modeling and Sentiment Analysis of Tweets about the Muslim Community in Spain since 2015," vol. 14, no. 6, p. 724, May,2023, https://doi.org/10.3390/rel14060724.

[13] A. Palanivinayagam, C. Z. El-Bayeh, and R. Damaševičius, "Twenty Years of Machine-Learning-Based Text Classification: A Systematic Review," vol. 16, no. 5, p. 236, April,2023, https://doi.org/10.3390/a16050236.

[14] X. Shu and Y. Ye, "Knowledge Discovery: Methods from data mining and machine learning," vol. 110, p. 102817, February,2023, https://doi.org/10.1016/j.ssresearch.2022.102817.

[15] R. Ulgasesa, A. B. P. Negara, and T. Tursina, "Pengaruh Stemming Terhadap Performa Klasifikasi Sentimen Masyarakat Tentang Kebijakan New Normal," vol. 10, no. 3, p. 286, September,2022, https://doi.org/10.26418/justin.v10i3.53880.

[16] E. Dewi, "Islam Liberal di Indonesia (Pemikiran dan Pengaruhnya dalam Pemikiran Politik Islam di Indonesia)," vol. 2, no. 2, pp. 18–32, January,2018, https://doi.org/10.14710/jiip.v2i2.2119.

[17] K. Bustamam-Ahmad, "Contemporary Islamic Thought in Indonesian and Malay World: Islam Liberal, Islam Hadhari, and Islam Progresif," vol. 5, no. 1, p. 91, June,2011, https://doi.org/10.15642/JIIS.2011.5.1.91-129.

[18] C. T. Agustina, "Pergerakan jaringan islam liberal (jil) di indonesia tahun 2001-2005," vol. 4, p. 242059, September,2012. [Online]. Available: https://www.neliti.com/publications/242059/

[19] D. E. Cahyani and I. Patasik, "Performance comparison of TF-IDF and Word2Vec models for emotion text classification," vol. 10, no. 5, pp. 2780–2788, October,2021, https://doi.org/10.11591/eei.v10i5.3157.

[20] Z. Nussbaum, J. X. Morris, B. Duderstadt, and A. Mulyar. (2024) Nomic Embed: Training a Reproducible Long Context Text Embedder. https://doi.org/10.48550/ARXIV.2402.01613.

[21] J. Mutinda, W. Mwangi, and G. Okeyo, "Sentiment Analysis of Text Reviews Using Lexicon-Enhanced Bert Embedding (LeBERT) Model with Convolutional Neural Network," vol. 13, no. 3, p. 1445, 2023-01-21, https://doi.org/10.3390/app13031445.

[22] X. Zhang, N. Thakur, O. Ogundepo, E. Kamalloo, D. Alfonso-Hermelo, X. Li, Q. Liu, M. Rezagholizadeh, and J. Lin, "**MIRACL** : A Multilingual Retrieval Dataset Covering 18 Diverse Languages," vol. 11, pp. 1114–1131, September,2023, https://doi.org/10.1162/tacl_a_00595.

[23] H. Abdelmotaleb, C. Mcneile, and M. Wojtyś, "A comparative study of word embedding techniques for classification of star ratings," vol. 297, p. 129037, February,2026, https://doi.org/10.1016/j.eswa.2025.129037.

[This page is intentionally left blank.]