International Journal of Informatics and Computation (IJICOM)
Vol. 8, No. 1, 2026
ISSN: 2685-8711, E-ISSN: 2714-5263 DOI: 10.35842/ijicom

Semantic Similarity Analysis of Hadith Matn in Indonesian Ṣaḥīḥ al-Bukhārī Corpus Using IndoBERT and Cosine Similarity

M. Didik R. Wahyudi[1], Noorhaidi Hasan[2], Agung Fatwanto[3]

Abstract
Digital hadith corpora create opportunities for computational analysis of semantic relationships among hadith texts that convey similar meanings through different wording. However, lexical-based similarity methods often fail to identify semantic proximity when hadith matn contain paraphrases, structural variations, or thematic similarities with limited word overlap. This study analyzes semantic similarity in the Indonesian-translated Ṣaḥīḥ al-Bukhārī corpus using IndoBERT and cosine similarity. We apply a quantitative text mining approach that includes data preprocessing, sanad–matn separation, narrator-based subcorpus construction, contextual embedding generation with IndoBERT, pairwise cosine similarity calculation, and similarity score categorization. The analysis focuses on two narrator-based subcorpora: Abu Hurairah (936 matn) and Anas bin Malik (744 matn), resulting in 696,384 comparison pairs. The results show that 68.06% of Abu Hurairah's matn and 80.65% of Anas bin Malik's matn belong to the high-similarity category, indicating substantial thematic overlap between the two subcorpora. In contrast, low-similarity matn reveal more distinctive thematic characteristics. These findings demonstrate that IndoBERT effectively captures semantic relationships beyond literal word matching and can support exploratory analysis, thematic mapping, and meaning-based hadith retrieval in digital hadith corpora.

Keywords:
Semantic Similarity; IndoBERT; Cosine Similarity; Text Mining; Hadith Matn

1. Introduction

The development of Natural Language Processing (NLP) has encouraged the use of computational techniques to analyze large-scale text corpora. Texts that were previously read manually can now be processed as digital data to identify patterns, measure similarity, cluster themes, and develop information retrieval systems. In this context, text mining serves as an approach for extracting knowledge from collections of textual documents through computational processes at the word, sentence, and document levels [1], [2], [3]. Several recent studies published in IJICOM also show that NLP and text mining can be used for various digital text analysis purposes, such as opinion classification on the use of Artificial Intelligence and the development of NLP-based chatbots for information services [4], [5]. In the era of Generative AI, mapping semantic proximity in hadith corpora has become important because AI systems require a knowledge base that can be traced, examined, and linked back to textual sources. The semantic similarity approach can support meaning-based search and reduce dependence on keyword-based search alone.

Corresponding Author: M. Didik R. Wahyudi (m.didik@uin-suka.ac.id)
1 M. Didik R. Wahyudi, UIN Sunan Kalijaga, m.didik@uin-suka.ac.id
2 Noorhaidi Hasan, UIN Sunan Kalijaga, noorhaidi@uin-suka.ac.id
3. Agung Fatwanto, UIN Sunan Kalijaga, agung.fatwanto@uin-suka.ac.id

One important issue in text mining is how to measure semantic proximity between texts, especially when two documents express similar ideas using different wording.

Text similarity measurement was initially dominated by lexical-based approaches, such as term frequency-inverse document frequency, vector space model, and cosine similarity [3], [6]. These approaches are effective in detecting document proximity based on word overlap or term occurrence patterns. Lexical-feature-based text mining approaches are also still widely used in recent studies, for example, in sentiment analysis of digital application reviews through preprocessing, model training, and classification performance evaluation stages [7]. However, lexical-based approaches have limitations when the texts being compared contain variations in wording, synonymy, paraphrases, or differences in sentence structure. In such cases, two texts may convey closely related meanings even though they do not share a high degree of word overlap. This limitation has encouraged the use of embedding-based text representations that are able to capture semantic information in a more contextual manner.

One text corpus that is particularly relevant for NLP-based analysis is the hadith corpus. Hadith is a religious text with a distinctive structure consisting of sanad and matn. The sanad contains the chain of transmission, while the matn contains the substance or content of the narration. In the tradition of hadith studies, hadith analysis is conducted through sanad criticism and matn criticism to assess the quality, continuity of transmission, and conformity of the narration's content [8], [9], [10]. However, when hadith texts are available in digital form, they can also be treated as textual datasets that enable computational analysis. This approach is not intended to replace classical hadith methodology, but to provide a quantitative tool for mapping patterns, themes, and semantic proximity among hadith matn.

The main problem addressed in this study is that hadith matn often contain semantic proximity even when expressed in different wording. Two narrations may discuss the same theme, closely related doctrinal contexts, or similar narrative patterns without using identical word sequences. Therefore, an embedding-based approach is needed to capture semantic relationships that are not always visible through word matching. The development of the Transformer model through the self-attention mechanism [11], along with BERT as a pre-trained language model based on bidirectional contextual representation [12], has opened new opportunities for analyzing text similarity at a more semantic level. For Indonesian texts, IndoBERT is particularly relevant because it was developed to support Natural Language Understanding tasks in the Indonesian language [13]. Fine-tuned IndoBERT was also used to capture semantic context, negation, and contrast in Indonesian texts more effectively than word-frequency-based approaches [14].

Based on this background, this study is designed to answer three main questions. First, how IndoBERT and cosine similarity can be used to measure semantic similarity among hadith matn in the Indonesian-translated Ṣaḥīḥ al-Bukhārī corpus. Second, how the distribution patterns of semantic similarity scores appear between the hadith matn narrated by Abu Hurairah and Anas bin Malik. Third, how the measurement results can be interpreted as indicators of thematic proximity among narrations without treating them as claims of hadith validity or as final historical evidence.

The contribution of this study lies in the application of IndoBERT and cosine similarity to map semantic similarity among Indonesian hadith matn and to support the development of meaning-based hadith information retrieval systems.

2. Related Works

Text mining has become one of the important approaches in the processing and analysis of digital text corpora. Aggarwal and Zhai [1] describe text mining as a field that includes various techniques for extracting patterns and knowledge from textual data, including classification, clustering, information retrieval, and document similarity analysis. Feldman and Sanger [2] also emphasize that text mining enables unstructured text to be processed into more organized and computationally analyzable information. Meanwhile, Manning et al. [3] provide an important foundation for information retrieval-based text processing, particularly in document representation, term weighting, and relevance measurement. These three works serve as a general foundation for this study because they show that text corpora can be analyzed as digital data to identify relationships between documents.

In text similarity measurement, lexical-based approaches have long been used through models such as term frequency-inverse document frequency, vector space model, and cosine similarity. Salton and McGill [6] serve as one of the key references in the development of modern information retrieval, particularly through document representation in vector space. In this approach, document proximity is calculated based on patterns of word or term occurrence. Manning et al. [3] further expand this discussion in the context of indexing, weighting, and document retrieval. This approach is relevant to text similarity research, but it has limitations because it is more sensitive to word similarity than to semantic proximity. Therefore, lexical-based approaches are not yet fully adequate for analyzing hadith matn, which often have variations in wording while still containing thematic proximity.

The development of the Transformer model has provided a new direction for text representation. Vaswani et al. [11] introduced the self-attention mechanism, which enables a model to capture relationships between tokens more flexibly within a text sequence. This approach later became the foundation for various modern language models, including BERT. Devlin et al. [12] developed BERT as a pre-trained language model capable of generating bidirectional contextual representations. This model has been widely used in various NLP tasks, such as text classification, question answering, information retrieval, and semantic similarity. In the context of this study, the development of Transformer and BERT is important because it enables text similarity analysis to rely not only on word overlap but also on more contextual meaning representations.

For Indonesian, IndoBERT has become one of the important models because it was specifically developed to support Natural Language Understanding tasks in Indonesian. Wilie et al. [13] introduced IndoNLU as a benchmark and resource for evaluating Indonesian language understanding, while also providing an IndoBERT model trained on a large-scale Indonesian corpus. In addition, Koto et al. [15] introduced IndoLEM as a benchmark for Indonesian NLP and IndoBERT as a BERT-based pre-trained language model. These studies show that IndoBERT achieves strong performance across various Indonesian language processing tasks. The relevance of IndoBERT in this study lies in its ability to represent Indonesian texts contextually. This is important because the corpus used in this study is the Indonesian translation of Ṣaḥīḥ al-Bukhārī. Therefore, IndoBERT was selected because it is more suitable than general models that were not specifically developed for Indonesian.

In hadith studies, computational approaches have been used for various purposes. Some studies focus on corpus construction, digitization, and the management of hadith data structures. Altammami et al. [16] constructed a bilingual hadith corpus using a segmentation tool that can separate hadith components, especially sanad and matn. Mahmood et al. [17] also developed a multilingual hadith dataset repository that can be used for text mining, information retrieval, and knowledge extraction research. The

contributions of these studies lie in providing databases and corpus structures that enable hadith to be processed digitally. This study is related to that direction because it uses a digital hadith corpus as the object of analysis, but its focus is not on dataset construction; rather, it focuses on measuring semantic similarity among hadith matn.

Other studies in hadith computing have used network analysis to map relationships among narrators in the sanad. Alam and Schneider [18] represented narrator chains in *Ṣaḥīḥ al-Bukhārī* as a social graph to identify influential narrators and patterns of hadith transmission. Saeed et al. [19] also used social network analysis to map hadith narrator networks and identify relational structures within transmission networks. These studies show that sanad structures can be analyzed computationally as relational networks. Although relevant to the development of data-driven hadith studies, these approaches focus more on sanad structure than on the substance of the matn. This study takes a different direction by focusing on hadith matn as textual units to be compared semantically.

In addition to corpus construction and sanad analysis, several studies have applied information retrieval techniques, latent semantic analysis, TF-IDF, and cosine similarity in hadith search systems or hadith text similarity measurement. Amrizal [20] applied TF-IDF and cosine similarity in a web-based information retrieval system for identifying hadith commentary. Darmalaksana et al. [21] developed a hadith search engine using latent semantic analysis and cosine similarity. Meanwhile, Yunus et al. [22] used cosine similarity and the Boyer-Moore method to detect hadith similarity in the context of women's fiqh. These studies demonstrate that hadith can be analyzed using text mining approaches. However, most of them still rely on lexical approaches, text retrieval, or representation models that do not yet fully capture deeper semantic context.

Based on these studies, previous computational hadith research has mostly focused on dataset construction, sanad analysis, and lexical similarity-based search systems. Studies that specifically use IndoBERT to map semantic similarity among Indonesian-translated hadith matn remain relatively limited.

3. Proposed Method

This study uses a quantitative method with a text mining approach and semantic similarity techniques based on IndoBERT and cosine similarity. This method is used to measure semantic similarity among hadith matn in the Indonesian-translated corpus. The selection of this method is based on the main objective of the study, namely, to conduct Natural Language Processing (NLP) analysis to map semantic proximity between texts. This mapping of semantic proximity is not intended to verify hadith authenticity, but to explore patterns of semantic similarity in a digital hadith corpus.

To support this objective, this study uses contextual text representations to map semantic proximity among hadith matn. IndoBERT was selected because BERT-based language models generate contextual embeddings that are more suitable for capturing semantic relationships beyond lexical overlap [12], [13], [23]. The selection of IndoBERT is also based on its suitability for the corpus used in this study, namely the Indonesian translation of Ṣaḥīḥ al-Bukhārī. In contrast, lexical representations such as TF-IDF rely on term occurrence and word overlap, making them less effective when two texts convey similar meanings using different wording [3], [6], [24]. Cosine similarity is used because it provides a simple and widely adopted measure for comparing vector representations in semantic similarity measurement [3], [6], [23]. In this study, cosine similarity was selected because the purpose of the analysis is not to perform classification, but to measure the

directional proximity between embedding vectors that represent hadith matn. Although IndoBERT and cosine similarity are commonly used approaches, their combination is considered relevant because it is stable, easily replicable, and suitable for exploring semantic proximity among hadith matn in Indonesian translation.

Basically, IndoBERT follows the basic principles of BERT, which is built on the Transformer encoder [12]. Therefore, the mathematical formulation used to explain the contextual representation mechanism of IndoBERT can refer to the self-attention mechanism in the Transformer architecture. The scaled dot-product attention mechanism is formulated as follows [11].

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (1)$$

In this equation, Q, K, and V are matrix representations formed from the input token sequence. The operation $\mathrm{QK^T}$ is used to calculate the degree of association among tokens in a sequence, while $\sqrt{\mathrm{d_k}}$ functions as a scaling factor to keep the dot product values stable. The softmax function produces attention weights for the relevant tokens, and these weights are then multiplied by V to generate contextual representations. Through this mechanism, each token in a hadith matn can be represented by considering the context of other tokens in the sentence.

After each hadith matn is represented as a semantic vector, the proximity between vectors is calculated using cosine similarity. This measure is used to determine how close the directions of two vectors are in the embedding space [3], [6], [23]. Mathematically, cosine similarity is formulated as follows:

$$sim(v_i, v_j) = \frac{v_i \,.v_j}{\|v_i\| \, \|v_j\|} \quad (2)$$

In this equation, $\mathrm{v_i}$ and $\mathrm{v_j}$ represent the semantic representation vectors of the two hadith matn being compared. The symbol $v_i.v_j$ denotes the dot product between the two vectors, while $\|v_i\|$ and $\|v_j\|$ denote the norm or length of each vector. A similarity value closer to 1 indicates that the two matn have stronger semantic proximity. Since each matn in one subcorpus is compared with all matn in the other subcorpus, this study takes the highest similarity score as the most semantically similar pair:

$$s_i = max_j sim(v_i, v_j) \quad (3)$$

In this equation, s_i indicates the highest similarity score for the i-th matn, while j represents all candidate comparison matn from the other subcorpus.

In this study, hadith is treated as a text corpus, while hadith matn serves as the main unit of analysis. The sanad is still used in the initial stage to identify narrators and construct subcorpora, but the similarity calculation is performed only on the matn text. This separation is important to ensure that the similarity scores are not influenced by repeated narrator names, transmission formulas, or relatively repetitive sanad structures. The conceptual workflow of the research method is shown in Fig. 1

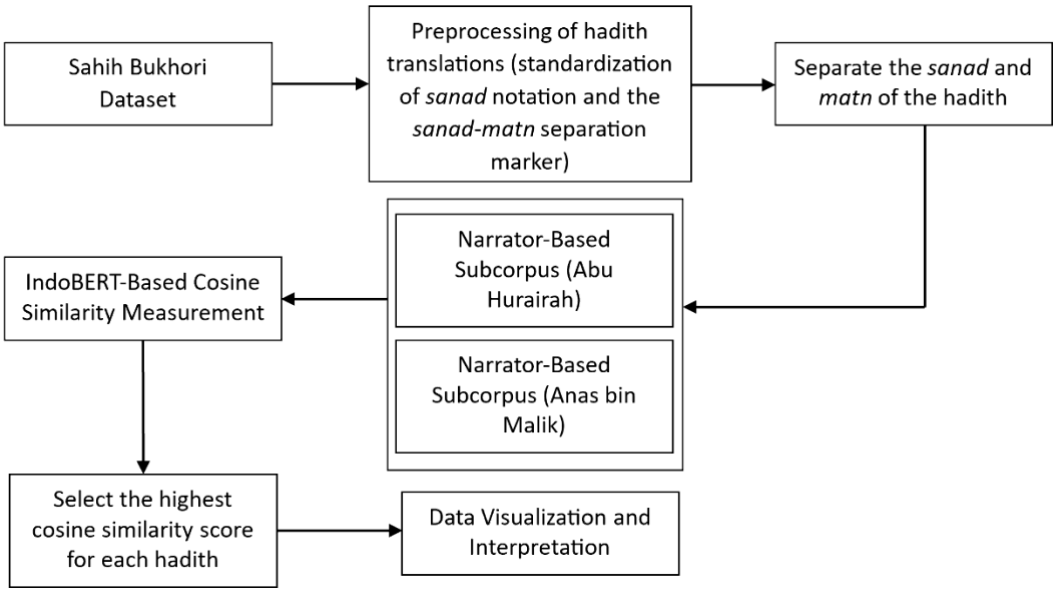

Fig. 1. Proposed method workflow

Based on Fig. 1, the study begins with the use of the *Ṣaḥīḥ al-Bukhārī* dataset in digital format as the main corpus. The dataset first undergoes a preprocessing stage to prepare the text for computational processing. This stage includes text normalization, standardization of sanad notation, uniform formatting of sanad-matn boundary markers, and removal of non-essential characters.

After preprocessing, the sanad and matn of each hadith are separated. The sanad is used to identify narrators and construct subcorpora, while the matn is used as the main unit of analysis. This separation is carried out to ensure that the similarity measurement is not influenced by repeated narrator names or transmission formulas in the sanad.

The next stage is the construction of two narrator-based subcorpora, namely the Abu Hurairah subcorpus and the Anas bin Malik subcorpus. Each matn in the two subcorpora is then processed through IndoBERT-based cosine similarity measurement. At this stage, IndoBERT is used to generate vector representations of the matn, while cosine similarity is used to calculate semantic proximity between vectors. From the comparison results, the highest cosine similarity score for each matn is selected as the most semantically similar pair. The final results are then visualized and interpreted to examine patterns of semantic proximity among hadith matn. The interpretation is conducted as a computational analysis of semantic proximity, not as a claim of hadith authenticity or historical transmission relationship.

4. Experimental Setup

The experiment for measuring semantic similarity among hadith matn was conducted using the Indonesian-translated corpus of *Ṣaḥīḥ al-Bukhārī* as the main data source. The experimental setup includes dataset preparation, narrator-based subcorpus construction, IndoBERT-based embedding representation, cosine similarity calculation, and score distribution analysis.

4.1 Data Collection

The data used in this study were obtained from the Hadith Data Sets repository [25]. The dataset consists of the Indonesian-translated corpus of *Ṣaḥīḥ al-Bukhārī*, which contains 7,008 hadith entries. Each data entry includes several attributes, namely the hadith number, book or chapter information, Arabic hadith text, and Indonesian translation. The *Ṣaḥīḥ al-Bukhārī* corpus was selected because it has a relatively consistent data structure and a sufficient number of entries for text mining analysis. In this experiment, the text used as the object of analysis is the Indonesian translation of the hadith because the language representation model used is IndoBERT. General information about the dataset used in the experiment is presented in Table 1.

Table 1. Dataset Description

Attribute	Description
Dataset	Indonesian translation of *Ṣaḥīḥ al-Bukhārī*
Source	*Hadith Data Sets* repository
Initial records	7,008 hadith
Text used	Indonesian translation
Unit of analysis	Hadith matan
Compared subcorpora	Abu Hurairah and Anas bin Malik
Main task	Semantic similarity measurement

Before being used in the experiment, the dataset was examined to ensure that the Indonesian translation column could be processed as text input. Empty records, duplicated records, or entries without adequate Indonesian translation were excluded from the analysis. Narrator names were also normalized to ensure consistency in subcorpus construction.

4.2 Experimental Procedure

This study conducted the experiment in a Python programming environment by first loading the dataset into a dataframe to support data cleaning, text separation, subcorpus construction, and similarity computation. We used the Indonesian translation column as the primary input and applied light preprocessing focused on technical cleaning and normalization rather than aggressive stopword removal or stemming, since Transformer-based models such as IndoBERT rely on contextual representations in which function words, word order, and sentence context contribute to semantic understanding. Following the principles of text mining, we performed data cleaning as an initial preparation stage before computational text representation and modeling. We then tokenized each hadith matn using the IndoBERT tokenizer and generated contextual embeddings through the IndoBERT model. To obtain sentence-level representations, we applied mean pooling to the token embeddings from the final hidden layer, resulting in a single semantic vector for each matn. Finally, this study calculated pairwise cosine similarity between the Abu Hurairah and Anas bin Malik subcorpora by comparing each matn in one subcorpus with every matn in the other and storing the results in a similarity matrix, which served as the basis for identifying semantically similar hadith matn pairs.

4.3 Data Analysis

This study analyzed the experimental results using descriptive and interpretative approaches. We organized the cosine similarity scores into a similarity matrix to examine semantic relationships between hadith matn narrated by Abu Hurairah and Anas bin Malik.

For each matn, we selected the highest cosine similarity score to represent its strongest semantic match in the other subcorpus. We then analyzed the distribution of these maximum scores and categorized them into low, medium, and high similarity levels to facilitate interpretation. In addition, this study qualitatively examined selected matn pairs with high and low similarity scores to assess whether the embedding-based similarities reflected meaningful semantic and thematic relationships in the texts.

5. Result and Analysis

Semantic similarity analysis among hadith matn was conducted using IndoBERT and cosine similarity on the Indonesian-translated corpus of *Ṣaḥīḥ al-Bukhārī*. The analysis results include data preprocessing, subcorpus construction, similarity score calculation, distribution of similarity levels, and thematic interpretation of matn with low and high similarity levels.

5.1 Data Preprocessing Result

The preprocessing stage produced a hadith corpus that was ready for computational processing. The initial dataset contained 7,008 hadith entries from *Ṣaḥīḥ al-Bukhārī*. After technical cleaning and light normalization, the dataset had a sufficiently consistent structure for sanad and matn separation. The resulting new dataset contains additional columns for sanad and matn, as shown in Fig. 2.

id1	terjemah	nontash	sanada	matna	id2	sanadI	matnI	sanad_akhir
1	Telah menceritakan kepada kami [Al Humaidi Abd...	حدثنا الحميدي عبد الله بن الزبير قال حدثنا سفي...	حدثنا الحميدي عبد الله بن الزبير قال حدثنا سفي...	يقول إنما الأعمال بالنيات وإنما لكل امرئ ما نو...	1	[Al Humaidi Abdullah bin Az Zubair], [Sufyan],...	["Semua perbuatan tergantung niatnya, dan (ba...	Umar_bin_Al_Khaththab
2	Telah menceritakan kepada kami [Abdullah bin Y...	حدثنا عبد الله بن يوسف قال أخبرنا مالك عن هشام...	حدثنا عبد الله بن يوسف قال أخبرنا مالك عن هشام...	فقال يا رسول الله كيف يأتيك الوحي فقال رسول ال...	2	[Abdullah bin Yusuf], [Malik], [Hisyam bin 'Ur...	["Wahai Rasulullah, bagaimana caranya wahyu t...	Aisyah
3	Telah menceritakan kepada kami [Yahya bin Buka...	حدثنا يحيى بن بكير قال حدثنا الليث عن عقيل عن ...	حدثنا يحيى بن بكير قال حدثنا الليث عن عقيل عن ...	من الوحي الرؤيا الصالحة في النوم فكان لا يرى ر...	3	[Yahya bin Bukair], [Al Laits], ['Uqail], [Ibn...	["Permulaaan wahyu yang datang kepada Rasulul...	Mamar
...	...	...	...	...	...	...	...	...
7006	Telah menceritakan kepada kami [Ali] telah men...	حدثنا علي حدثنا هشام أخبرنا معمر عن الزهري ح و...	حدثنا علي حدثنا هشام أخبرنا معمر عن الزهري ح و...	يكون حقا قال فقال النبي صلى الله عليه وسلم ثل...	7006	[Ali], [Hisyam], [Ma'mar], [Azzuhri], [Ahmad b...	["Mereka tidak ada apa-apanya." Para sahabat ...	Aisyah
7007	Telah menceritakan kepada kami [Abu Nu'man] te...	حدثنا أبو النعمان حدثنا مهدي بن ميمون سمعت محم...	حدثنا أبو النعمان حدثنا مهدي بن ميمون سمعت محم...	قال يخرج ناس من قبل المشرق ويقرءون القرآن لا ي...	7007	[Abu Nu'man], [Mahdi bin maimun], [Muhammad bi...	["Akan muncul beberapa orang dari arah timur,...	Abu_Said_Al_Khudzri
7008	Telah menceritakan kepadaku [Ahmad bin Isykab]...	حدثني أحمد بن إشكاب حدثنا محمد بن فضيل عن عمار...	حدثني أحمد بن إشكاب حدثنا محمد بن فضيل عن عمار...	كلمتان حبيبتان إلى الرحمن خفيفتان على اللسان ث...	7008	[Ahmad bin Isykab], [Muhammad bin Fudlail], ['...	["Ada dua kalimat yang disukai Ar Rahman, rin...	Abu_Hurairah

7008 rows × 9 columns

Fig. 2. Example of the new dataset containing sanad and matn columns

5.2 Text Processing and Modeling Result

The text processing and modeling stage includes sanad and matn separation, narrator-based subcorpus construction, vector representation using IndoBERT, and similarity calculation among matn using cosine similarity. The first process in this stage is the separation of sanad and matn. The sanad contains the chain of transmission, while the matn contains the content or substance of the hadith. This separation is necessary to ensure that the embedding process is not influenced by repeated narrator names, transmission formulas, or relatively repetitive sanad structures. The extracted hadith matn are used in the similarity calculation, while the sanad is used to identify narrators and construct subcorpora.

After the separation process is completed, a new dataset is created with additional sanad and matn columns. The sanad column is used for narrator identification, while the matn column is used as input for embedding construction. From the corpus processing results, 917 narrator names were identified in the *Ṣaḥīḥ al-Bukhārī* dataset used in this study. A summary of the number of narrators in the dataset is shown in Fig. 3.

	Perawi	Jumlah
0	Abu_Hurairah	936
1	Anas_bin_Malik	744
2	Aisyah	740
3	Ibnu_Abbas	485
4	Ibnu_Umar	351
...	...	...
913	Ubadah_bin_shamit	1
914	Ubbay_bin_Kab_Al_anshari	1
915	Barra_bin_Azib	1
916	Khaitsamah	1
917	Amru_bin_Taghlib	1

918 rows × 2 columns

	Perawi	Jumlah
0	Abu_Hurairah	936
1	Anas_bin_Malik	744
2	Aisyah	740
3	Ibnu_Abbas	485
4	Ibnu_Umar	351
5	Abdullah	221
6	Abdullah_bin_Umar	191
7	Jabir_bin_Abdullah	115
8	Abu_Musa	89
9	Jabir	88

Fig. 3. Number of narrators in the Ṣaḥīḥ al-Bukhārī dataset

This study selected two subcorpora, namely hadith narrated by Abu Hurairah (936 matn) and Anas bin Malik (744 matn), because both narrators contributed a large number of narrations to the dataset. We performed a comprehensive pairwise comparison by comparing each matn in the Abu Hurairah subcorpus with every matn in the Anas bin Malik subcorpus, resulting in 696,384 comparison pairs. This approach ensured that the similarity analysis was not influenced by sample selection, as every matn had the opportunity to be compared with all matn in the other subcorpus. We calculated semantic similarity by generating embedding vectors with IndoBERT and measuring their cosine similarity. Higher cosine similarity values indicate greater semantic or thematic proximity between two matn rather than direct textual similarity. The resulting output consisted of hadith pairs containing the identities of both hadith, their matn texts, and the corresponding cosine similarity scores, as illustrated in Fig.

similarity_abm_ah-ALL-ed2

id	idabm	mtabm	idah	mtah	scoresim
0	12	tidaklah berima	8	iman memiliki le	0,42570746
0	12	tidaklah berima	13	maka zat jiwaku	0,78651226
0	12	tidaklah berima	25	iman allah rasul	0,55120003
0	12	tidaklah berima	32	tanda-tanda mu	0,57894754
0	12	tidaklah berima	34	barangsiapa me	0,5399576
0	12	tidaklah berima	35	aku mendengar	0,61433876
0	12	tidaklah berima	36	barangsiapa me	0,6590489

Record: 1 of 696384 No Filter Search

Fig. 4. Result of cosine similarity calculation using IndoBERT

From the pairwise comparison process, 696,384 combinations of similarity scores were obtained based on the Abu Hurairah and Anas bin Malik subcorpora. These scores served as the basis for determining the matn pairs with the highest semantic proximity and for constructing the distribution of similarity categories in the next stage.

5.3 Similarity Score Distribution

The similarity scores generated in this study were then grouped into three main categories. Scores below 0.60 were categorized as low similarity. Scores from 0.60 to less than 0.70 were categorized as medium similarity. Scores of 0.70 and above were categorized as high similarity. For a more detailed analysis, the high-similarity category was further divided into three ranges: 0.70 to less than 0.80, 0.80 to less than 0.90, and 0.90 and above. This categorization follows the approach used in near-duplicate detection, which applies similarity thresholds to distinguish levels of textual proximity [26].

The cosine similarity calculation for hadith matn narrated by Abu Hurairah shows that most matn have pairs with a high level of similarity to hadith matn narrated by Anas bin Malik. The score distribution is presented in Table 2.

Table 2. Similarity Score Distribution of Abu Hurairah's Hadith Matan

Similarity score	Frequency	Percentage
Similarity score < 0.60	36	3.85%
0.60 ≤ similarity score < 0.70	263	28.10%
0.70 ≤ similarity score < 0.80	475	50.75%
0.80 ≤ similarity score < 0.90	147	15.71%
Similarity score ≥ 0.90	15	1.60%
Total	**936**	**100%**

Based on Table 2, the hadith narrated by Abu Hurairah tends to show semantic similarity with the hadith narrated by Anas bin Malik. If the high-similarity category is defined as a similarity score of ≥ 0.70, this category covers most of the data. Under this scheme, high similarity reaches 68.06%, while medium and low similarity account for smaller proportions. The visualization of the similarity distribution in the Abu Hurairah subcorpus is shown in Fig. 5.

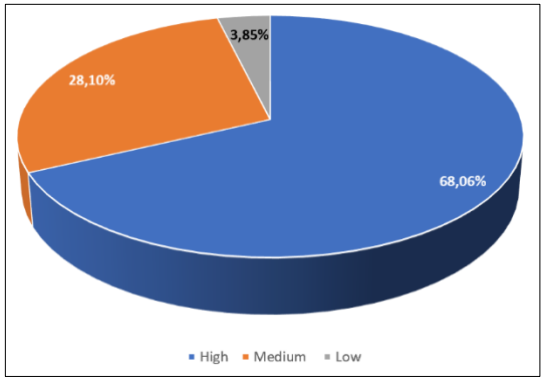

Fig. 5. Percentage of similarity categories in Abu Hurairah's hadith dataset

Based on Fig. 5, the low-similarity category in the Abu Hurairah subcorpus contains 36 matn (3.85% of the total). The dominant themes in this group include charity, emancipation of slaves, pledged animals, the story of Prophet Isa, paradise and hell, treatment with habbatus sauda, anger control, wealth and poverty, marriage, signs of the Day of Judgment, and hypocrisy. These findings suggest that several matn narrated by Abu Hurairah discuss themes that have limited semantic correspondence with the matn in the Anas bin Malik subcorpus. We then conducted the analysis in the opposite direction by examining the similarity-score distribution of hadith matn narrated by Anas bin Malik against those narrated by Abu Hurairah. The results are presented in Table 3.

Table 3. Similarity Score Distribution of Anas bin Malik's Hadith Matan

Similarity score	Frequency	Percentage
Similarity score < 0.60	17	2.28%
0.60 ≤ similarity score < 0.70	127	17.07%
0.70 ≤ similarity score < 0.80	348	46.77%
0.80 ≤ similarity score < 0.90	221	29.70%
Similarity score ≥ 0.90	31	4.17%
Total	**744**	**100%**

Table 3 shows that the hadith narrated by Anas bin Malik also tends to have semantic similarity with the hadith narrated by Abu Hurairah. The high-similarity category reaches 80.65%, while medium similarity accounts for 17.07% and low similarity for 2.28%. The visualization of the similarity distribution in the Anas bin Malik subcorpus is shown in Fig. 6.

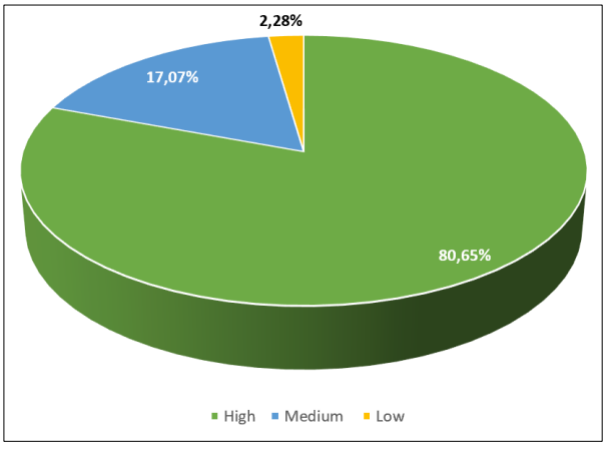

Fig. 6. Percentage of similarity categories in Anas bin Malik's hadith dataset

Based on Fig. 6, the low-similarity group in the Anas bin Malik hadith dataset consists of 17 matn, or 2.28%. The dominant themes in this group include images, food for the Prophet, the Prophet's marriage to Shafiyyah bint Huyay ibn Akhtab, belonging to a community, the death of a child, treatment with cupping and gaharu, and provisions for the afterlife. The percentage of high similarity in the Anas bin Malik subcorpus is higher than that in the Abu Hurairah subcorpus. Computationally, this difference may occur because the Abu Hurairah subcorpus used as the comparison set is larger, thereby providing a broader space for semantic matching. The larger the number of comparison candidates, the greater the possibility that a matn will find a pair with a high similarity score.

5.4 Thematic Demonstration of Similarity Score Ranges

To clarify how cosine similarity scores represent the level of semantic proximity among matn, this study presents several examples of matn pairs from different score ranges. These examples were selected to show that similarity values do not function merely as quantitative scores, but can also be used to examine variations in thematic proximity. Lower scores indicate clearer differences in theme or context, while higher scores indicate stronger proximity in wording, theme, or meaning structure. Examples of matn pairs across several similarity ranges are presented in Table 4.

Table 4. Examples of Hadith Matan Pairs across Similarity Score Ranges

ID Anas bin malik	Matan Anas bin Malik	ID Abu Hurairah	Matan Abu Hurairah	Score
3786	Nabi shallallahu alaihi wasallam mendoakan kebinasaan kaum telah membunuh sahabat beliau birul maunah selama tiga puluh hari beliau mendoakan kebinasaan ril lahyan ushayyah...	2079	cambuklah kemudian dia berzina cambuklah kemudian juallah melakukan ketiga keempat kalinya	0,513
6310	...diantara tanda kiamat adalah- ilmu diangkat kebodohan merajalela khamer ditenggak zina mewabah jumlah laki-laki menyusut jumlah wanita melimpah ruah hingga ada lima puluh wanita berbanding seorang laki-laki	5931	kami hafal abu az zinad al araj abu hurairah periwayatan berkata allah memiliki sembilan puluh sembilan nama seratus kurang satu tidaklah seseorang menghafalnya ia masuk surga…	0,595
683	luruskanlah shaf-shaf kalian sesungguhnya aku melihat kalian balik punggungku dan orang kami merapatkan bahunya bahu temannya kakinya kaki temannya	2990	seseorang kalian selalu dihitung berada dalam shalat shalat mengekangnya (orang tersebut menanti shalat ditegakkan) malaikat mendoakan ya allah ampunilah dan rahmatilah selama belum berdiri tempat shalatnya telah berhadats	0,655
5753	orang memuji allah maka aku mendoakannya yang tidak memuji allah	2727	aku diperintahkan memerangi manusia hingga mengucapkan laa ilaaha illallah (tidak ilah allah) maka barang siapa mengucapkan laa ilaaha illallah sungguh telah terlindung jiwa hartanya dariku…..	0,700

ID Anas bin malik	Matan Anas bin Malik	ID Abu Hurairah	Matan Abu Hurairah	Score
5402	apakah nabi shallallahu alaihi wasallam pernah shalat menggunakan sandal dia menjawab ya pernah	1144	nabi shallallahu alaihi wasallam melarang seseorang shalat bertolak pinggang	0,764
3627	...nabi shallallahu alaihi wasallam tiba madinah diantara shahabat beliau paling tua abu bakr kemudian menyemirnya menggunakan daun inai katam daun pewarna lainnya hingga warna rambutnya nampak kemerah-merahan	4232	....beliau mengucapkan doa sambil berdiri sujud, ya allah selamatkanlah ayyasy bin abu rabi ah salamah bin hisyam al walid bin al walid orang-orang lemah kalangan kaum mukmin ya allah timpakan siksaan-mu bani mudhar jadikanlah tahun-tahun seperti tahun-tahun yusuf	0,800
691	sesungguhnya imam dijadikannya imam untuk diikuti jika takbir bertakbirlah kalian rukuk rukuklahkalian ia mengangkat kepala angkatlah kepala kalian ia mengucapkan samiallahu liman hamidah semoga allah mendengar orang memuji-nya ucapkanlah kalian rabbanaa lakal hamdu ya rabb kami milik engkaulah segala pujian jika ia sujud sujudlah kalian	680	dijadikannya imam untuk diikuti janganlah kalian menyelisihnya jika rukuk rukuklah kalian mengucapkan samiallahu liman hamidah ucapkanlah rabbanaa lakal hamdu jika ia sujud sujudlah kalian ia shalat duduk shalatlah kalian semuanya duduk luruskanlah shaf lurusnya shaf merupakan bagian sempurnanya shalat	0,852
3012	sesungguhnya surga sebuah pohon jika pengendara berjalan bawah naungannya seratus tahun lamanya akan melewatinya	3013	sesungguhnya surga sebuah pohon bayangannya ditempuh para pengendara memerlukan waktu seratus tahun lamanya bacalah firman allah jika kamu mau artinya ujung panah seseorang kalian surga lebih baik tempat matahari terbit terbenam	0,897
1577	kendarailah unta itu orang menjawab unta untuk qurban maka beliau shallallahualaihiwasallam mengulangi perintahnya kendarailah unta itu tiga kali	2550	kendarailah unta itu orang menjawab wahai rasulullah unta untuk qurban maka beliau shallallahu alaihi wasallam mengulangi kendarailah unta itu celaka kamu ini perintah	0,938

ID Anas bin malik	Matan Anas bin Malik	ID Abu Hurairah	Matan Abu Hurairah	Score
			beliau kedua kalinya yang ketiga	
379	...maimun bin siyah bertanya anas bin malik wahai abu hamzah apa menjadikan haramnya darah harta seorang hamba ali menjawab siapa bersaksi laa ilaaha illallah tidak tuhan berhak disembah allah menghadap kiblat kita shalat sepeti shalat kita dan memakan sembelihan kita dia muslim baginya hak dan kewajiban seorang muslim	2727	aku diperintahkan memerangi manusia hingga mengucapkan laa ilaaha illallah. maka barang siapa mengucapkan laa ilaaha illallah sungguh telah terlindung jiwa hartanya dariku dengan haqnya perhitunganya allah	0,992

Table 4 shows that each score range has a different semantic character. At low scores, such as 0.513 and 0.595, the matn pairs show fairly clear thematic differences. Matn concerning the Prophet's supplication regarding the Bi'r Ma'unah incident, the punishment for adultery, the signs of the end times, and the names of Allah belong to different thematic contexts; therefore, their semantic proximity is relatively low. In the medium range, such as the score of 0.655, the matn pairs begin to show thematic overlap, for example, both being related to prayer practices, although their discussion contexts are not entirely the same.

In the high-score range, such as 0.700 to 0.897, thematic proximity appears stronger. The example of matn pairs concerning prayer while wearing sandals and the prohibition of praying with hands on the waist indicates proximity within the theme of prayer practices. Other pairs concerning the imam in prayer and the tree in Paradise also show stronger similarities in theme and wording patterns. Meanwhile, very high scores, such as 0.938 and 0.992, indicate very strong semantic proximity because the two matn share nearly the same theme, meaning structure, and partial wording. These examples demonstrate that IndoBERT and cosine similarity are able to distinguish different levels of proximity among matn, ranging from pairs with weak thematic relations to pairs with very strong similarities in meaning and wording.

5.5 Discussion

The dominance of high-similarity scores indicates substantial semantic overlap between the Abu Hurairah and Anas bin Malik subcorpora in the embedding space generated by IndoBERT. This finding suggests that the combination of IndoBERT and cosine similarity effectively captures thematic relationships among hadith matn, even when their wording differs. Such results may partly reflect the characteristics of the hadith corpus, which contains recurring religious themes, similar teaching patterns, and repeated linguistic structures. The use of Indonesian translations may further increase textual homogeneity through consistent translation choices. Consequently, similarity scores should be

interpreted as indicators of semantic proximity rather than evidence of historical proximity or shared transmission events. While high scores reflect thematic coherence between the two subcorpora, low scores reveal relatively distinctive themes and semantic differentiation. Methodologically, these findings demonstrate the usefulness of IndoBERT and cosine similarity as exploratory tools for hadith corpus analysis, enabling researchers to identify semantically related matn, detect distinctive themes, and map meaning relationships across texts. The approach also has practical implications for information systems, including semantic hadith retrieval, thematic recommendation, and text clustering. More broadly, this study highlights the potential of Indonesian language models for semantic analysis of translated religious texts and provides a foundation for developing semantic search, Retrieval-Augmented Generation (RAG), and explainable AI applications that require traceable links between generated outputs and original hadith sources.

6. Conclusion

This study investigated semantic similarity among hadith matn in the Indonesian-translated Ṣaḥīḥ al-Bukhārī corpus using IndoBERT and cosine similarity. To overcome the limitations of lexical-based approaches, we applied a text mining framework consisting of data preprocessing, sanad–matn separation, narrator-based subcorpus construction, contextual embedding generation, and similarity analysis. Using a dataset of 7,008 hadith, we focused on two narrator-based subcorpora: Abu Hurairah (936 matn) and Anas bin Malik (744 matn). Pairwise comparisons generated 696,384 matn pairs, and the resulting similarity scores revealed substantial semantic overlap between the two subcorpora. The high-similarity category accounted for 68.06% of the Abu Hurairah subcorpus and 80.65% of the Anas bin Malik subcorpus, indicating that IndoBERT embeddings can effectively capture thematic relationships beyond literal textual similarity. At the same time, low-similarity scores highlighted distinctive legal, ethical, eschatological, and event-specific themes within each subcorpus.

The findings should be interpreted as computational indicators of semantic and thematic proximity rather than evidence of hadith authenticity, narrator relationships, or historical transmission. Therefore, embedding-based similarity is best viewed as an exploratory tool for mapping meaning relationships among hadith matn in a digital corpus and should be complemented by qualitative and domain-specific hadith analysis. This study contributes to the application of Natural Language Processing in Indonesian religious texts and demonstrates the potential of IndoBERT for semantic analysis of hadith corpora. Future research may expand the corpus coverage, incorporate expert validation, integrate the original Arabic texts, and compare IndoBERT with alternative approaches such as TF-IDF, other embedding models, and different similarity metrics to provide a more comprehensive evaluation of semantic similarity measurement in hadith studies.

Acknowledgment

The authors would like to express their sincere gratitude to Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM), UIN Sunan Kalijaga Yogyakarta, for providing financial support for the publication of this article. This support has contributed significantly to the completion and dissemination of this research.

References

[1] C. C. Aggarwal and C. X. Zhai, *Mining text data*, vol. 9781461432. in Mining Text Data, vol. 9781461432. Springer US, 2013, p. 522. doi: 10.1007/978-1-4614-3223-4.

[2] R. Feldman and J. Sanger, *The Text Mining Handbook*. in The Text Mining Handbook. Cambridge University Press, 2009, p. 424. doi: 10.1017/cbo9780511546914.

[3] P. R. Christopher D. Manning Hinrich Schutze, *Introduction to Information Retrieval*. Cambridge, United Kingdom: Cambridge University Press, 2008, p. 506.

[4] R. Melinda Yuniar and S. Mulyati, "Sentiment Classification of Student Opinions on AI Utilization Using Naive Bayes Algorithm," *ijicom*, vol. 7, no. 1, pp. 279–290, Jun. 2025, doi: 10.35842/ijicom.v7i1.122.

[5] B. Ismarizal, R. Pradila, H. Hendrian, and H. A. Ramadhan, "Leveraging Chatbot Model for Tourism Information Services using NLP and ANN," *ijicom*, vol. 8, no. 1, pp. 59–68, Feb. 2026, doi: 10.35842/ijicom.v8i1.170.

[6] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[7] F. M. Julianto, A. Turmudi Zy, and E. Rilvani, "Sentiment Analysis on Canva Reviews Using Naive Bayes Method," *ijicom*, vol. 7, no. 1, pp. 86–98, Feb. 2025, doi: 10.35842/ijicom.v7i1.107.

[8] K. Amin, *Menguji Kembali Keakuratan Metode Kritik Hadis*. Hikmah, 2009.

[9] J. A. C. Brown, *The canonization of al-Bukhārī and Muslim: the formation and function of the Sunnī Ḥadīth canon*. in Islamic History and Civilization, no. 69. Leiden: Brill, 2007.

[10] A. M. Yaqub, *Kritik Hadis*. Jakarta: Pustaka Firdaus, 1995.

[11] A. Vaswani *et al.*, "Attention Is All You Need," 2017, *arXiv*. doi: 10.48550/ARXIV.1706.03762.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018, *arXiv*. doi: 10.48550/ARXIV.1810.04805.

[13] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," 2020, *arXiv*. doi: 10.48550/ARXIV.2009.05387.

[14] E. R. S. H. Saputra, A. C. Frobenius, and R. F. A. Aziza, "Evaluating Public Trust in the Animation Industry: A Comparative Sentiment Analysis Using Random Forest and Fine-Tuned IndoBERT," *ijicom*, vol. 8, no. 1, pp. 30–39, Jan. 2026, doi: 10.35842/ijicom.v8i1.168.

[15] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[16] S. Altammami, E. Atwell, and A. Alsalka, "Constructing a Bilingual Hadith Corpus Using a Segmentation Tool," *Proc. 12th Language Resources and Evaluation Conf. (LREC 2020)*, pp. 3390–3398, May 2020.

[17] A. Mahmood, H. Ullah, F. K., M. Ramzan, and M. Ilyas, "A Multilingual Datasets Repository of the Hadith Content," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 2, 2018, doi: 10.14569/IJACSA.2018.090224.

[18] T. Alam and J. Schneider, "Social Network Analysis of Hadith Narrators from Sahih Bukhari," in *2020 7th International Conference on Behavioural and Social Computing (BESC)*, Bournemouth, United Kingdom: IEEE, Nov. 2020, pp. 1–5. doi: 10.1109/BESC51023.2020.9348299.

[19] S. Saeed, S. Yousuf, F. Khan, and Q. Rajput, “Social network analysis of Hadith narrators,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, Part B, pp. 3766–3774, Jun. 2022, doi: 10.1016/j.jksuci.2021.01.019.

[20] V. Amrizal, “Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) dan Cosine Similarity pada Sistem Temu Kembali Informasi untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Hadits Shahih Bukhari-Muslim),” *J. Teknik inform.*, vol. 11, no. 2, pp. 149–164, Nov. 2018, doi: 10.15408/jti.v11i2.8623.

[21] W. Darmalaksana, C. Slamet, W. B. Zulfikar, I. F. Fadillah, D. S. Maylawati, and H. Ali, “Latent semantic analysis and cosine similarity for hadith search engine,” *TELKOMNIKA*, vol. 18, no. 1, p. 217, Feb. 2020, doi: 10.12928/telkomnika.v18i1.14874.

[22] B. M Yunus, “Similarity Detection for Hadith of Fiqh of Women using Cosine Similarity and Boyer Moore Method,” *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 1, pp. 65–73, Feb. 2020, doi: 10.30534/ijatcse/2020/11912020.

[23] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” Aug. 27, 2019, *arXiv*: arXiv:1908.10084. doi: 10.48550/arXiv.1908.10084.

[24] C. De Boom, S. Van Canneyt, S. Bohez, T. Demeester, and B. Dhoedt, “Learning Semantic Similarity for Very Short Texts,” in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, Atlantic City, NJ, USA: IEEE, Nov. 2015, pp. 1229–1234. doi: 10.1109/ICDMW.2015.86.

[25] A. Kamar, “Hadith-Data-Sets,” GitHub repository. Accessed: Dec. 23, 2023. [Online]. Available: https://github.com/abdelrahmaan/Hadith-Data-Sets

[26] H. Hajishirzi, W. Yih, and A. Kolcz, “Adaptive near-duplicate detection via similarity learning,” in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, Geneva Switzerland: ACM, Jul. 2010, pp. 419–426. doi: 10.1145/1835449.1835520.