Accredited by National Journal Accreditation (ARJUNA) Managed by
Ministry of Research, Technology, and Higher Education, Republic Indonesia with Second Grade (Peringkat 2, Sinta 2)
since year 2017 to 2021 according to the decree No. 10/E/KPT/2019.

Published online on the journal's webpage: **http://jurnal.iaii.or.id**

RESTI JOURNAL
(System Engineering and Information Technology)
Vol. 4 No. 3 (2020) 551 - 557 ISSN Electronic Media: 2580-0760

Improving Document Retrieval with Spelling Correction for Weak and Fabricated Indonesian-Translated Hadith

Muhammad Zaky Ramadhan[1], Kemas Muslim Lhaksmana[2]
[1,2]Informatics, School of Computing, Telkom University
[1]m.zakyramadhan1924@gmail.com, [2]kemasmuslim@telkomuniversity.ac.id

Abstract

Hadith has several levels of authenticity, among which are weak (dhaif), and fabricated (maudhu) hadith that may not originate from the prophet Muhammad PBUH, and thus should not be considered in concluding an Islamic law (sharia). However, many such hadiths have been commonly confused as authentic hadiths among ordinary Muslims. To easily distinguish such hadiths, this paper proposes a method to check the authenticity of a hadith by comparing them with a collection of fabricated hadiths in Indonesian. The proposed method applies the vector space model and also performs spelling correction using symspell to check whether the use of spelling check can improve the accuracy of hadith retrieval, because it has never been done in previous works and typos are common on Indonesian-translated hadiths on the Web and social media raw text. The experiment result shows that the use of spell checking improves the mean average precision and recall to become 81% (from 73%) and 89% (from 80%), respectively. Therefore, the improvement in accuracy by implementing spelling correction make the hadith retrieval system more feasible and encouraged to be implemented in future works because it can correct typos that are common in the raw text on the Internet.

Keywords: hadith, vector space model, symspell, spelling correction, document retrieval.

1. Introduction

Hadith is a primary source of Islamic law after the Qur'an. As a primary source of law means every activity carried out must be in accordance with what is regulated in the Qur'an and hadith. But the existence of hadith that originated from the Prophet Muhammad was tainted by the appearance of weak and fabricated hadith where the origin and interpretation are used in various motives [1]. Believing and practicing *sharia* based on a weak or even fabricated hadith is an act that is condemned by Rasulullah SAW with the threat to take his seat in hell. Weak and fabricated hadiths with good or bad intentions remain counted as an act of counterfeiting so that there should be no action based solely on weak and fabricated hadith [2].

Weak and fabricated hadith is the lowest and the worst level of authenticity compared to other hadith. Fabricated hadith are hadith that are intentionally made or falsified while weak ones are hadith that are the result of unintentional forgery, occurring because there are errors in the chain of narrators of the hadith [3]. Examples of hadith that are often encountered lately are "*Perselisihan di antara umatku adalah rahmat*" (The difference of opinions between my people is a mercy for my people), whereas according to *syarah* Al-Albani this hadith has no source and because of that statement, many Muslims after the time of the Imams especially today continue to disagree and debate in many ways that involve a creed and practice, the statement is also contrary to the Qur'an Al-Anfal verse 46 "*Dan taatlah kepada Allah dan Rasul-Nya dan janganlah kamu berbantah-bantahan, yang menyebablan kamu menjadi gentar dan hilang kekuatanrnu…*" (And obey Allah and His Messenger, and do not dispute and thus lose courage and then your strength would depart, etc.). The verse makes it clear that disputes and animosity are not from God. Al-Albani also said in his *syarah* [4] that these words will negatively impact Muslims from time to time. Disputes caused by differences between *mazhab* have reached its climax, even the fanatical followers of the *mazhab* are not reluctant to disbelieve and branded the followers of other *mazhab* as heretics.

Currently, the only method of checking back the authenticity of a hadith is by asking a religious expert or by searching the hadith manually on physical books and electronic documents. Therefore, we need a system to search weak and fabricated hadith as a media to double-check whether the information obtained is based on

Accepted by editor: 28-04-2020 | Final Revision: 24-05-2020 | Online Publication: 20-06-2020

weak and fabricated hadith and save ourselves from the threat of hellfire and religious debates resulting from taking weak and fabricated hadith as the basis of our action.

In recent years, the implementation of natural language processing in the Qur'an and hadith are used to conduct information retrieval [5]. Natural language processing can process and analyze information on documents so that it can be used to search whether a piece of text information that is said to be a hadith can be used as a query as a search parameter for a collection of Indonesian translated documents of weak and fabricated hadith.

There are several studies to conduct information retrieval. Research conducted by Vibkhav et al. [6] was conducted to analyze and explain how the vector space model works. Hanum et al. [7] conduct research on information retrieval in the form of halal product queries from Malay language documents using the latent semantic indexing method and generate an accuracy value of 86%. Jbara et al. [8] researched the classification of texts that existed in 1321 *Sahih al Bukhari hadith* by classifying classes that have the same topic then the testing is done by comparing with the similarity coefficient table and generate an accuracy of 73%. Humaini et al. [9] design an algorithm for Indonesian translated *Qur'an* based on the vector space model and TF-IDF which generate 98.70% accuracy.

In previous studies, the use of vector space models has the highest level of accuracy, which is 98.70%. The second best method uses latent semantic indexing with 86% accuracy and then the similarity coefficient table method with 73% accuracy. However, the three related studies use different data sets, research with latent semantic indexing using Malay language data sets, while vector space models and coefficient tables use Indonesian translated data sets. Therefore, the vector space model was chosen because the dataset used is Indonesian translated and is the best method for conducting retrieval documents [10].

Based on related research, this paper proposes to build a weak and fabricated hadith document retrieval system using vector space models and can correct spelling by using the spell checking method Symspell to calculate the increase in accuracy of document retrieval, because the test data used will be in the form of raw text from the internet that contains spelling errors. Vector space models work by representing text information in documents in a frequency index vector [6]. The spell checking method that will be used as a functional additive to improve the accuracy of document retrieval is the symmetric deletion spell checking algorithm or the Symspell method because Symspell can provide faster spelling correction than conventional methods like Peter Norvig [11].

2. Research Method

This research has several steps starting from the acquisition of corpus data, test data and training data, dictionary development, normalization using vector space model, spell correction using Symmetric Delete Spelling Correction or Symspell, feature extraction using TF-IDF and searching the relevant document using cosine similarity that can be described as in Figure 1.

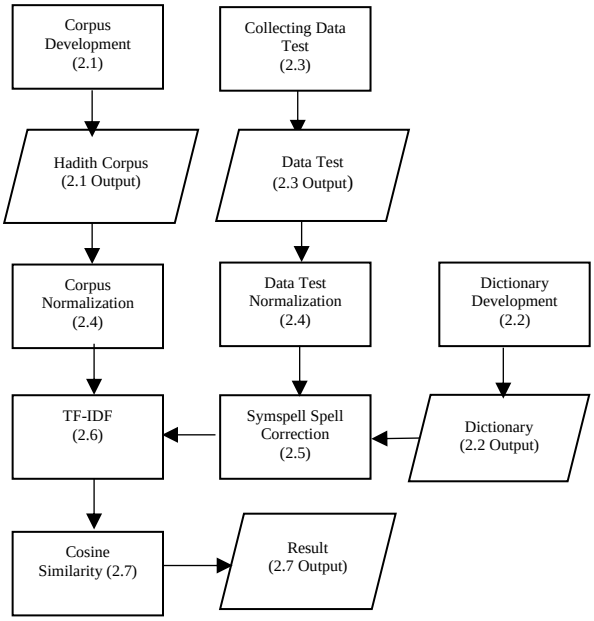

Figure 1. Research Methodology

2.1. Corpus Development

The development of weak and fabricated hadith corpus is taken from a book [4]. Hadith's data amounts to 2000 data collected manually in digital .csv format. The book was chosen because it is the most comprehensive book about weak and fake hadith which is translated into Indonesian and contains *syarah* which explains why the hadith classified as weak or fabricated hadith in detail. To ensure the validity and reliability of the data, the collection process to .csv form is carried out using optical character recognition and rechecked manually.

2.2. Dictionary Development

The development of the dictionary was taken from Wikipedia articles dump file, news portals dump file, and Islamic corpus, for a total of 108.902 sentences from Leipzig Indonesian corpus. Dictionary development is done by calculating the number of times the word appears on Indonesian documents and ranks the appearance of words from words that most often appear to words that rarely appear in the form of bigram and unigram.

2.3. Collecting Data Test

The collection of test data was taken by searching in Indonesian social media, forums, and news portals based

on the book [12]. Data test amounts to 85 hadith with at least one spelling mistake.

2.4. Data Normalisation

Data normalization is the process of converting a text sentence into a Bag-of-Words format. Bag-of-words is the accumulation of word tokens that are derived from the processing of sentences by carrying out a series of processes, including: Case folding, filtering, stemming, tokenization [13]. The text normalization process can be explained as follows:

1. Case Folding, this Stage changes all text to lowercase.
2. Filtering, this stage eliminates symbols in the text such as ". (dot)", ", (comma)" and empty words that have no meaning, such as: "*dan*", "*di*", "*yang*", and so on.
3. Stemming, this stage returns the words to their basic form by removing prefix, suffix, infixes, and confixes.
4. Tokenization, stages of cutting sentences into separate words using spaces as a delimiter. For example, sentence. "*puasa malam hari*" becomes "*puasa*", "*malam*", "*hari*".

2.5. Symspell Spelling Correction

Symspell algorithm was created by Wolf Garbe in 2012 based on Peter Norvig's spell checking research with faster performance because of Symspell algorithm only uses delete operation [11].

Stages performed by the Symspell algorithm can be explained as follows:

1. Delete letters that are in the dictionary and save the results.
2. Load the results of stage number 1
3. For each word that is detected typo, delete a letter from the word.
4. Find each deletion result in the dataset
5. If found, add the word to the correct word recommendation.

For example, if the word "*hari*" is in the dictionary and the word "*yari*" is a word that a typo is detected. The two words delete operations are performed which can be seen in Table 1. below:

Table 1. Symspell Deletion Example

Dictionary	Typo
Ari	**Ari**
Hai	Yai
Har	Yar

In Table 1. there is a similarity with the word "Ari", so the word "Hari" will enter the recommendation list of the correct word.

Validation of the spelling correction by calculating the total correctness of hits and false positives using the equation 1 and 2 [14] and uses data test from Indonesian social media and news portals.

$$Correction\ hit = \frac{\sum typos\ corrected}{\sum typos}\ x\ 100\% \quad (1)$$

$$False\ Positive = \frac{\sum false\ positive}{\sum correct\ words}\ x\ 100\% \quad (2)$$

Correction hits are the number of spelling errors corrected divided by the number of spelling errors, while false positive is the number of words that are already correct but autocorrect is performed divided by the number of correct words.

2.6. Feature Extraction

The success of the system in retrieving documents from the entered query can be influenced by the weighting scheme used for both local and global coverage [15]. The weighting that will be used is the TF-IDF method. The term frequency is the frequency of the appearance of a term or word in a document. The greater the number of occurrences of a word in a document, the greater the similarity in value. The term frequency calculation used in this research is a raw calculation, i.e. the term frequency value is calculated based on the number of words found in the document. For example, there is the word "*puasa*" twice in the document, then the term frequency value of "*puasa*" is two.

IDF calculates how the terms are widely distributed in existing document collections and shows the relationship between the availability of a term in all documents. The smaller the number of documents containing the term, the higher the IDF value will be. IDF value calculations can be done with the following equation 3.

$$IDF = log\ (D/Df) \quad (3)$$

Information:
IDF = Inverse of the frequency value in the document
D = Total number of documents
Df = Number of documents containing the term

TF-IDF calculates the weight of the relationship of a term with a document by calculating the frequency of occurrence of the term in the document and calculating how the term is widely distributed in the existing document collection [15].

$$W = TF * IDF \quad (4)$$

Information:
W = Document weight
TF = Term frequency value
IDF = Inverse of the frequency value in the document

2.7 Document Retrieval

The vector space model is a method for performing information retrieval by observing the distance or similarity of terms. The searched documents are seen as vectors that have distance and direction. In the vector space model, the relevance level of the query for the searched document is based on the proximity of the document vector and the query vector [16].

Suppose there are *n* different words in the dictionary. These words will form a vector space that has dimensions as large as n. Each word *i* in a document or query receives the weight *W* obtained from equation 4. The vector matrix that will be formed with *D* is the sentence of the document and *T* is the term or the word is as follows:

$$\begin{bmatrix} & T1 & \cdots & Tt \\ D1 & W11 & \cdots & Wt1 \\ \cdots & \cdots & \cdots & \cdots \\ Dn & W1n & \cdots & Wtn \end{bmatrix}$$

To calculates the size of the angular similarity between the document vector and the query vector we will use cosine similarity using equation 5 with *Q* is the input query [15].

$$Cosine(D,Q) = \frac{D.Q}{||D|| \ . \ ||Q||} \quad (5)$$

To validate the retrieval system, we need to calculate the mean average precision, average precision, and recall using the equation 6, 7, and 8.

$$MAP = \frac{1}{|Q|} . \sum_{i=1}^{Q} AP(Q_i) \quad (6)$$

$$AP = \sum_{i=1}^{N} \left(\frac{TP_{-\mathrm{i}}}{TPrank_{-i}} \right) \quad (7)$$

$$Recalls = \frac{|(Relevantdocument) \cap (Output)|}{(Relevantdocument)} \quad (8)$$

Mean Average Precision is a score obtained from measuring system performance in a series of queries [16]. MAP represents the average value of average precision from all queries.

Q is the set of queries. *AP* (*Qi*) is the average precision value to query *i*. The average precision value for each query will be calculated first. Average precision is a measure of precision by calculating the sequence of relevant documents at the output of the system based on rank [17]. For example in Table 2., assumes there are 3 relevant documents for a query and our system return 5 most relevant output result.

Table 2. Precision Calculation Example

Output rank-	True/False	Precision	Information
1	True	1/1	Relevant on rank-1
2	False	-	Not relevant
3	False	-	Not relevant
4	True	2/4	Relevant on rank-4
5	True	3/5	Relevant on rank-5

From Table 2. above we can calculate the AP score of queries: $\frac{\frac{1}{1}+\frac{2}{4}+\frac{3}{5}}{3} = 0.53$

Recalls are the number of relevant documents obtained compared to all relevant documents. Recall value indicates the system's ability to search the relevant documents. The maximum value of recall is 1 and the minimum value is 0. Recall valued at 1 means that the system successfully outputs all the relevant documents while 0 means that none of the outputs are relevant.

3. Result and Discussion

The unauthentic hadith search system that was created was tested using data obtained from social networks where each data has a typing error. Each data is entered as a query and two tests are performed, the test applying spelling correction and the test without applying spelling correction.

3.1. Dataset

Corpus hadith can be represented in Table 3. The first column contains the Id of the hadith that will later be used to calculate the precision of the system, the second column is the content of weak and fabricated hadiths in Indonesian, and the third column is the *syarah* or argument of why the hadith is declared weak or fabricated.

Table 3. Corpus Hadith Example

Id	Content	Syarah
SDMJ 1N001	*Agama adalah akal, siapa saja…* (Religion is intelect, anyone, etc.)	*Hadis tersebut batil…* (The hadith is vanity, etc.)
SDMJ 1N002	*Barang siapa shalatnya…* (Whoever prays, etc.)	*Walaupun hadits…* (Despite the hadith, etc.)

The example of test data that will be used as a query in the system is "*Perselisihannan di antara ummatku adalah rahmat*" (The difference of opinions between my people is a mercy for my people) with the words "*Perselisihannan*" and "*ummatku*" are the misspelled words that need to be corrected.

Table 4. Experiment Result of All Spelling Mistakes

	ADD	DEL	SUB	SUB,DEL	SUB,ADD	SEGMENT	SEGMENT,ADD	TRANS
Hit	36	48	37	2	0	4	0	4
Not Hit	19	11	16	0	1	0	1	0
Total Typo	55	59	53	2	1	4	1	4
Correction Rate	0.65	0.81	0.70	1.00	0.00	1.00	0.00	1.00

Table 5. Correction Failure Example

Input	Gold Standard	Suggestion
Mencpi	*Mencari* (Seek)	*Men cpi*
Teruserang	*Terus terang* (Honestly)	*Terserang* (Attacked)

3.2. Spelling Correction Test

Spelling mistakes that will be tested:

1. SUB
   Substitute a letter in the word with another letter. For example, in the word "*idabah*" the letter "*d*" in the word will be replaced by the letter "*b*" so that the word becomes "*ibadah*".
2. SUB dan DEL
   Substitute and remove letters in the word. For example, in the word "*ibadshh*" the letter "*s*" in the word will be replaced by the letter "*a*" and will erase the last letter "*h*" so the word becomes "*ibadah*".
3. SUB dan ADD
   Substitute and add letters in the word. For example, in the word "*iadshh*" the letter "*s*" in the word will be replaced by the letter "*a*" and "*b*" letter will be added so the word becomes "*ibadah*".
4. SEGMENT
   Delete or add a space to the word. For example, the word "*ibadahmalam*" space will be added so the word becomes "*ibadah malam*".
5. SEGMENT dan ADD
   Add letters and delete or add spaces to the word. For example, in the word "*ibadahmala*" space and "*m*" letter will be added so the word becomes "*ibadah malam*".
6. TRANSPOSE
   Swap the position of the letters in the word. For example, in the word "*ibaadh*" letter "a" and "d" the position of the letters will be swapped so that the word becomes "*ibadah*".

Table 4. shows the accuracy of all tested spelling mistakes. From a total of 179 misspelled data, 131 were successfully corrected, resulting in a total correction rate of 73%. Spelling errors SUB, ADD and SEGMENT, ADD which cannot provide the expected spelling suggestions that can be seen in Table 5 below:

3.3. Searching Validation

The selected documents are pre-processed similarly to the training dataset. The average precision and recall used are 5, average precision@5 means calculating the precision of the 5 documents obtained when considering the document ranking position, while MAP is the average of average precision@5 for all queries. Recall@5 means how exactly 5 documents were obtained compared to the gold standard [18]. The results of the accuracy of the searches performed can be seen in Table 6. below:

Table 6. System Precision and Recall Result

	MAP	Average Recall@5
With Spelling Correction	81%	89%
Without Spelling Correction	73%	80%

3.4. Searching Application

The user interface is built using the Tkinter Library with a screen as shown in Figure 2., with the user first entering a weak or fabricated hadith as a query to be searched and then pressing the search button.

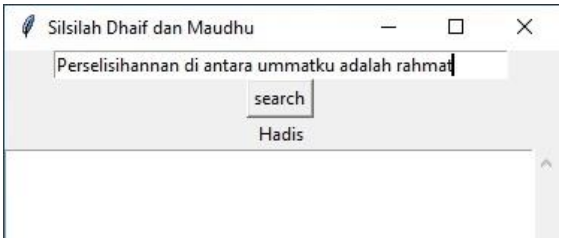

Figure 2. "*Perselisihannan di antara ummatku adalah rahmat*" (The Difference of Opinions between my People is a Mercy for My People) Used as Search Input in the Application

Figure 3. displays the hadith found in the first text area and the *syarah* in the second text area. The results of the searching for the hadith in Figure 2. were found and have *syarah* that the hadith "*Tidak ada sumbernya. Para pakar hadits telah berusaha mendapatkan sumbernya dengan meneliti dan menelusuri sanadnya, namun tidak menemukannya…"* (has no source. The hadith experts

have tried to obtain its source by researching and searching for its *sanad*, but did not find it, etc.).

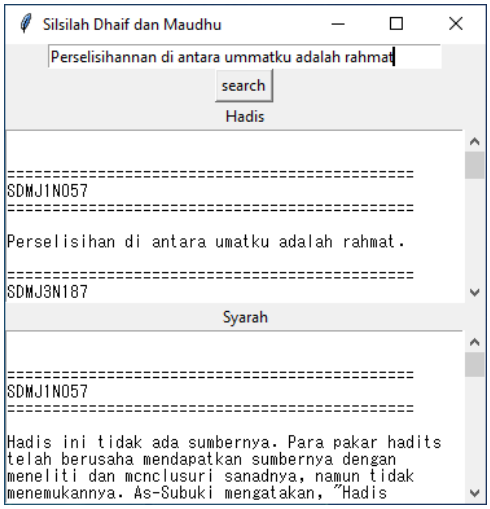

Figure 3. The Hadith that was Searched for was Found and Displayed along with its *Syarah* on the Application

3.5. Discussion

In contrast to previous related works, the accuracy may not be as high as in previous related works which are 98.70% compared to 89% recall in this paper. But, testing in this paper is done using raw text while in previous related works, the testing of the retrieval system uses a few words from the document text itself that have no spelling errors, ignoring the fact that the text on the Internet is a raw text that often has spelling errors. Therefore, the resulting accuracy of previous related works does not represent the reality of using a search system.

The final result of combining the document retrieval process with spelling check can be seen in Table 6. that searching using spelling correction can increase the mean average precision score by 8% and average recall by 9%. The increase occurred because, for example, the words "*Perselisihannan*" and "*ummahku*" in the feature extraction stage did not find any documents that have the same word, so that the words that have weight are only "*rahmat*" because the words "*di*", "*antara*", and "*adalah*" will be filtered. By using spelling correction, the word "*Perselisihannan*" can be corrected to "*Perselisihan*" and the word "*ummatku*" can be corrected to "*umatku*", so that the weight of the resulting document is greater and results in better precision and recall.

Based on the result, the spelling correction system also has a weakness in finding words that are names of people like “Juraji”, it will be considered a misspelling and will be corrected to "Juraij".

4. Conclusion

The use of autocorrect modules in this study can detect and give correct suggestions of misspellings that produce 81% mean average precision and 89% average recall greater 8% and 9% than without autocorrect that produces a 73% mean average precision and 80% average recall. Increased precision and recall scores occur because, without the application of the spelling correction Symspell in searching documents, there will be words that are misspelled that do not give weight to the relevant documents. The ability to correct misspelled words makes the system created in this paper more feasible to check a hadith authenticity to help distinguish weak and fabricated hadith among ordinary Muslims.

The autocorrect model that was made still has the disadvantage of correcting words that are the names and consists of multiple operations. By using named entity recognition, it should be able to recognize name entities that should not be evaluated, thus the models can give better accuracies. Furthermore, by using more data for training our models, we believe that they will perform better.

References

[1] R. Aslamiah, “Hadis Maudhu dan Akibatnya,” *Al-Hiwar J. Ilmu dan Tek. Dakwah*, vol. 4, no. 6, pp. 24–34, 2017, doi: 10.18592/al-hiwar.v4i6.1214.

[2] A. G. Fawwaz, “The Fabrication of Hadith,” University of Jordan, 2018.

[3] A. H. Usman and R. Wazir, “The Fabricated Hadith: Islamic Ethics and Guidelines of Hadith Dispersion in Social Media,” *Turkish Online J. Des. Art Commun.*, vol. Special Ed, pp. 804–808, 2018, doi: 10.7456/1080sse/114.

[4] M. N. Al-Albani, *Silsilah hadits dha’if dan maudhu’*. Jakarta: Gema Insani Press, 2005.

[5] E. Atwell, C. Brierley, K. Dukes, M. Sawalha, and A. Sharaf, “An Artificial Intelligence approach to Arabic and Islamic content on the internet,” in *Proc NITS’2011 National Information Technology Symposium, King Saud University, Saudi Arabia. Data protection statements*, 2011, no. 1, pp. 1–8, doi: 10.13140/2.1.2425.9528.

[6] V. K. Singh and V. K. Singh, “Vector Space Model: an Information Retrieval System,” in *Proceedings of BITCON-2015 Innovations For National Development*, 2015, pp. 141–143.

[7] H. M. Hanum, Z. A. Bakar, N. A. Rahman, M. M. Rosli, and N. Musa, “Using Topic Analysis for Querying Halal Information on Malay Documents,” *Procedia - Soc. Behav. Sci.*, vol. 121, no. 19, pp. 214–222, 2014, doi: 10.1016/j.sbspro.2014.01.1122.

[8] K. Jbara, “Knowledge Discovery in Al-Hadith Using Text Classification Algorithm,” *J. Am. Sci.*, vol. 6, no. 11, pp. 409–419, 2010.

[9] I. Humaini, L. Wulandari, D. Ikasari, and T. Yusnitasari, “Penerapan Algoritma Tf-Idf Vector Space Model (Vsm) Pada Information Retrieval Terjemahan Al Quran Surat 1 Samai Dengan Surat 16 Berdasarkan Kesamaan Makna Implementation of TF-IDF Vector Space Model (VSM) Algorithm in Information Retrieval of AL QURA,” 2019, pp. 525–534.

[10] M. A. Saloot, N. Idris, R. Mahmud, S. Ja’afar, D. Thorleuchter, and A. Gani, “Hadith data mining and classification: a comparative analysis,” *Artif. Intell. Rev.*, vol. 46, no. 1, pp. 113–128, 2016, doi: 10.1007/s10462-016-9458-x.

[11] W. Garbe, “1000X Faster Spelling Correction.” 2017, Accessed: Feb. 16, 2020. [Online]. Available: https://towardsdatascience.com/symspellcompound-10ec8f467c9b.

[12] A. Sabiq, *Hadits lemah dan palsu yang popular di Indonesia.* Gresik: Pustaka Al-Furqan, 2009.

[13] S. Vijayarani, J. Ilamathi, and M. Nithya, "Preprocessing Techniques for Text Mining," *Int. J. Comput. Sci. Commun. Networks*, vol. 5, no. 1, pp. 7–16, 2015.

[14] V. Christanti Mawardi, N. Susanto, and D. Santun Naga, "Spelling Correction for Text Documents in Bahasa Indonesia Using Finite State Automata and Levinshtein Distance Method," in *MATEC Web of Conferences*, 2018, pp. 1–16, doi: 10.1051/matecconf/201816401047.

[15] T. Sabbah *et al.*, "Modified frequency-based term weighting schemes for text classification," *Appl. Soft Comput. J.*, vol. 58, pp. 193–206, 2017, doi: 10.1016/j.asoc.2017.04.069.

[16] A. Aziz, R. Saptono, and K. P. Suryajaya, "Implementasi Vector Space Model dalam Pembangkitan Frequently Asked Questions Otomatis dan Solusi yang Relevan untuk Keluhan Pelanggan," *Sci. J. Informatics*, vol. 2, no. 2, pp. 111–121, 2015, doi: 10.15294/sji.v2i2.5076.

[17] Y. Rochmawati and R. Kusumaningrum, "Studi Perbandingan Algoritma Pencarian String dalam Metode Approximate String Matching untuk Identifikasi Kesalahan Pengetikan Teks," *J. Buana Inform.*, vol. 7, no. 2, pp. 125–134, 2016, doi: 10.24002/jbi.v7i2.491.

[18] I. R. Ponilan, Adiwijaya, M. A. Bijaksana, and A. S. Raharusun, "Search relevant retrieval on indonesian translation hadith document using query expansion and smoothing probabilistic model," in *Journal of Physics: Conference Series*, 2019, pp. 1–12, doi: 10.1088/1742-6596/1192/1/012032.